U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

INSTITUTO DE INVESTIGAÇÃO
E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Study of human regulatory mutations associated with pancreatic diseases by humanizing the zebrafish genome.

Inês Maria Batista da Costa

Mestrado em Biologia Celular e Molecular

Departamento de Biologia da Faculdade de Ciências da Universidade do Porto (FCUP)

Orientador

José Bessa, PhD, Instituto de Investigação e Inovação em Saúde (i3S)

Coorientador

Chiara Perrod, PhD, Instituto de Investigação e Inovação em Saúde (i3S)

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No 680156 – ZPR).

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto,

_____/_____/_____

Study of human regulatory mutations associated with pancreatic diseases by humanizing the zebrafish genome.

Inês Maria Batista da Costa

Mestrado em Biologia Celular e Molecular

Departamento de Biologia da Faculdade de Ciências da Universidade do Porto (FCUP)

Orientador

José Bessa, PhD, Instituto de Investigação e Inovação em Saúde (i3S)

Coorientador

Chiara Perrod, PhD, Instituto de Investigação e Inovação em Saúde (i3S)

Inês Maria Batista da Costa, BSc

Mestrado em Biologia Celular e Molecular

Departamento de Biologia

Faculdade de Ciências da Universidade do Porto (FCUP)

Rua do Campo Alegre s/n

4169-007 Porto, Portugal

up201405724@fc.up.pt / ines.costa@i3s.up.pt

Tel.: +351 917457361

Supervisor

José Bessa, PhD

Vertebrate Development and Regeneration Research Group

Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto

Rua Alfredo Allen, 208

4200-135 Porto, Portugal

jose.bessa@ibmc.up.pt

Tel.: +351 925254475

Co-supervisor

Chiara Perrod, PhD

Vertebrate Development and Regeneration Research Group

Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto

Rua Alfredo Allen, 208

4200-135 Porto, Portugal

chiara.perrod@ibmc.up.pt

Eu, Inês Maria Batista da Costa, aluna com o número 201405724 do Mestrado em Biologia Celular e Molecular da edição de 2017/2018, declaro por minha honra que sou a autora da totalidade do texto apresentado, não apresento texto plagiado e tomei conhecimento das consequências de uma situação de plágio.

Inês Maria Batista da Costa

27 de setembro de 2019

Agradecimentos

Ao meu orientador, Doutor José Bessa, pela oportunidade de participar neste projeto e orientar o meu trabalho de forma a que conseguisse concretizar este objetivo. À minha coorientadora, Chiara Perrod, pela enorme disponibilidade que mostrou para me ajudar em todas as alturas, pelos ensinamentos e pela amizade – Grazie! À Ana Gali, por fazer o possível e o impossível para me acompanhar no início da minha experiência em laboratório. A todos os restantes membros do grupo VDR – Ana Eufrásio, Ana Gali, Fábio Ferreira, Isabel Guedes, Joana Marques, Joana Teixeira, João Amorim, Marta Duque e Vítor Silva – pelo companheirismo, amizade e apoio que foram essenciais para me dar força e manter-me focada nos pontos positivos da vida de um investigador. A todos os membros honorários do VDR que também contribuíram para esta experiência.

Aos meus pais, por me darem a oportunidade de estudar e por acreditarem nas minhas capacidades. Ao melhor irmão do mundo, que me acompanha desde sempre e continua a ser uma das minhas raízes. À Ana Bompastor, por estar sempre presente, pela confiança e pelo apoio enorme. Ao Reyzão, a melhor companhia de todas as horas. À minha madrinha, por celebrar como ninguém todas as minhas pequenas conquistas. A toda a minha família pelo apoio incondicional.

E por fim, aqueles agradecimentos especialmente especiais. À minha besuga, por partilhar comigo cinco anos incríveis, pela confiança e pela amizade (que ainda agora começou). Ao melhor grupo de amigos do mundo, que me conhecem melhor do que eu me conheço a mim, e nunca se cansam de me apoiar. Ao André (o meu segundo irmão), Di e Duda (as minhas manas), Gustavo, Helena, Inês, Raquel, Sara e Tana. Obrigada por me levantarem quando estou em baixo, e por me fazerem acreditar que consigo chegar a Marte quando já estão comigo no topo do mundo. Às minhas quase-médicas preferidas, Maria e Mg.

A todos os que tenho o orgulho de ter como amigos e com quem partilho a minha vida.

Dedico este trabalho e a etapa final deste percurso à minha maior estrela no céu. Espero que me estejas a ver agora, vó. Vou estar sempre contigo.

Resumo

O diabetes tipo 2 (DT2) é uma doença poligénica com prevalência mundial, que depende de fatores de risco genéticos, epigenéticos e ambientais. Estudos de associação genómica (GWAS) mostraram que polimorfismos de nucleótido único (SNPs) não-codificantes estão associados à suscetibilidade a DT2. Alguns desses SNPs não só se sobrepõem a regiões caracterizadas por marcas epigenéticas de elementos cis-reguladores (CREs), como também residem dentro de áreas regulatórias de genes pancreáticos, como é evidenciado por dados da técnica de Hi-C realizada em ilhotas pancreáticas humanas. Um desses SNPs situa-se no *locus* do gene "*Pancreatic and Duodenal Homeobox 1*" (*PDX1*).

O fator de transcrição PDX1 desempenha um papel crucial no desenvolvimento inicial do pâncreas, na diferenciação de linhagens endócrinas e na função das células beta. Mutações heterozigóticas estão associadas a diabetes de uma forma monogênica de diabetes de manifestação precoce (MODY4), enquanto níveis reduzidos da expressão de *PDX1* são frequentemente observados em DT2.

Neste trabalho, pretendemos introduzir o *locus PDX1* humano no genoma do peixe-zebra, fornecendo, portanto, um modelo animal humanizado para realizar estudos *in vivo*. Concluímos que um cromossoma artificial bacteriano (do inglês, BAC) de 200 kilo bases contendo o gene *PDX1* representa seu contexto regulatório transcricional. O PDX1-BAC foi manipulado de forma a inserir locais de reconhecimento da transposase *Tol2* para aumentar a eficiência da transgénese e um repórter GFP para gerar uma proteína PDX1 de fusão para acompanhar sua expressão *in vivo*. O elemento transponível PDX1-BAC foi injetado no peixe-zebra no estadio de uma célula para gerar linhas transgénicas, que podem ser facilmente manipuladas pela edição do genoma do CRISPR-Cas9, de forma a caracterizar *in vivo* o papel de polimorfismos conhecidos e novos CREs identificados. Além disso, ao explorar o *locus PDX1* no que diz respeito a marcas epigenéticas, acessibilidade e conformação da cromatina, selecionamos seis putativos CREs de *PDX1*. Além disso, validamos parte dessas sequências isoladas humanas para a função intensificadora ou isoladora *in vivo* em peixe-zebra.

Esta abordagem fornecerá um poderoso modelo *in vivo* para investigar elementos não codificantes que regulam os níveis de expressão de *PDX1*, elucidando os mecanismos moleculares que contribuem para a suscetibilidade de DT2. A avaliação do impacto *in vivo* de mutações humanas no desenvolvimento de características

associadas à diabetes pode refinar estratégias de diagnóstico de DT2 e intervenções terapêuticas.

Palavras-chave: Diabetes Tipo 2 (DT2), *Pancreatic and Duodenal Homeobox 1* (*PDX1*), Genome-Wide Association Studies (GWAS), polimorfismos de nucleótido único (SNPs), regulação em cis, intensificador, isolador, contexto regulatório, *Danio rerio*, transgénese, *Tol2*.

Abstract

Type-2 Diabetes (T2D) is a polygenic disease with a worldwide prevalence, which depends on genetic, epigenetic and environmental risk factors. Genome-Wide Association Studies (GWAS) have shown that non-coding single nucleotide polymorphisms (SNPs) are associated to T2D susceptibility. Some of these SNPs not only overlap with regions characterised by epigenetic marks of cis-regulatory elements (CREs), but also reside within the regulatory landscapes of pancreatic genes, as seen by Hi-C data for human pancreatic islets. One of such SNPs lays in the *locus* of the *Pancreatic and Duodenal Homeobox 1* (*PDX1*) gene.

The transcription factor PDX1 plays a crucial role in early pancreas development, differentiation of endocrine lineages and beta cell function. Heterozygous mutations are associated with Maturity-Onset Diabetes of the Young (MODY4), an early-onset monogenic form of diabetes, while reduced *PDX1* expression levels are often observed in T2D.

In this work we aim to introduce the human *PDX1 locus* into the zebrafish genome, therefore providing a humanized animal model to perform *in vivo* studies. We reasoned that a Bacterial Artificial Chromosome (BAC) spanning 200 kilobases containing the *PDX1* gene represents its transcriptional regulatory landscape. The PDX1-BAC was engineered by inserting *Tol2* transposase recognition sites to enhance transgenesis efficiency and a GFP reporter to generate a fusion PDX1 protein to trace its expression *in vivo*. The PDX1-BAC transposable element was injected in one-cell stage zebrafish to generate transgenic lines, which can be easily manipulated by CRISPR-Cas9 genome editing to characterise *in vivo* the role of known polymorphisms and novel identified CREs. Additionally, by screening the *PDX1 locus* for epigenetic marks, chromatin accessibility and conformation, we selected six putative *PDX1* CREs. We further validated part of these human isolated sequences for enhancer or insulator function *in vivo* in zebrafish.

This approach will provide a powerful *in vivo* model to investigate non-coding elements regulating *PDX1* expression levels, elucidating the molecular mechanisms contributing to T2D susceptibility. Assessing the impact of human mutations for the development of diabetes-associated traits *in vivo* could refine T2D diagnostic strategies and therapeutic interventions.

Keywords: Type-2 Diabetes (T2D), *Pancreatic and Duodenal Homeobox 1* (*PDX1*), Genome-Wide Association Studies (GWAS), Single Nucleotide Polymorphism (SNP), cis-regulation, enhancer, insulator, regulatory landscapes, *Danio rerio*, transgenesis, *Tol2*.

Table of Contents

Table Index

Figure Index

Abbreviations

Chromosome conformation capture (3C)

Ampicillin resistance gene (AmpR)

Bacterial artificial chromosome (BAC)

Base-pair (bp)

Cis-regulatory element (CRE)Chromatin Immunoprecipitation (ChIP)

Days post-fertilization (dpf)

Deoxyribonucleic acid (DNA)

Expression Quantitative Trait *Loci* (eQTLs)

Genome-Wide Association Studies (GWAS)

Green fluorescence protein (GFP)

Hours post-fertilization (hpf)

Kanamycin resistance gene (KanR)

Maturity-onset diabetes of the young (MODY)

Minute/minutes (min)

Pancreatic and duodenal homeobox 1 (PDX1)

Polymerase chain reaction (PCR)

Ribonucleic acid (RNA)

Room Temperature (RT)

Single-nucleotide polymorphism (SNP)

Transcription factor (TF)

Transcription factor binding site (TFBS)

Type-2 Diabetes (T2D)

Wild-type (wt)

Introduction

1) Chromatin structure and Epigenetics

Epigenetics refers to DNA modifications involved in modulation of gene function, through changes in chromatin structure and accessibility without alteration at DNA sequence level [1]. In eukaryotes, the genetic information required for normal cell function is contained within the nucleus. Double-stranded deoxyribonucleic acid (DNA) is packed together with proteins, histones, forming DNA-protein complexes called nucleosomes. Nucleosomes are the core units of chromatin, which comprise 146 bp of a DNA sequence wrapped around an octamer of histone proteins [2].

Chromatin organization is fundamental for cell function, meaning that specific chromatin three-dimensional (3D) structure is associated with cell division and gene expression. On one hand, throughout cell division, the chromatin is arranged in its highest compact form. Several nucleosomes assemble together in loops and establish the units of chromatin, forming chromatin fibers. Chromatin packaging allows to arrange the genetic material in a compact way inside the cell nucleus [2]. On the other hand, during interphase, specific genomic regions are characterised by layers of chromatin accessibility, ensuring proper gene expression. Chromatin accessibility is related to the availability of DNA sequences to the transcriptional machinery [3].

Chromatin comprehends regions that are highly condensed and coiled – termed heterochromatin –, and regions that are more accessible – euchromatin. The constitutive heterochromatin is composed by structural chromosomic regions that retain their original condensed state during the interphase of the cell cycle – these include the centromere and telomeres. The other type of heterochromatin is called facultative, accordingly to its adjustable state of condensation, that is mainly dependent on histone modifications and differs between cell types [2]. In euchromatin, as the structure is more relaxed and available to transcription factors (TFs), these regions of chromatin are mainly associated with gene expression and transcriptional regulation [4].

The frequency of nucleosomes is enriched in facultative and constitutive heterochromatin, while it is decreased in euchromatin holding DNA regulatory sequences. The enrichment on nucleosomes contributes to determine chromatin accessibility, and therefore to gene expression regulation [5]. The modulation of nucleosome affinity to the chromatin depends on the histones and their posttranslational modifications, constituting one of the key epigenetic mechanisms of transcriptional regulation, often referred as “histones code”. Histones are proteins that can undergo

covalent modifications upon translation, through methylation, phosphorylation, acetylation, ubiquitylation and sumoylation processes. Facultative heterochromatin and inactive gene promoters are characterised by the presence of histone H3 lysine 27 trimethylation (H3K27me3) mark that leads to gene repression and silencing. Additionally, constitutive heterochromatin, is predominantly marked by other repressive modifications – H3K9 di- and trimethylation (H3K9me2 and H3K9me3). Specific histone modifications are used in the prediction of regulatory elements, such as enhancers. Putative enhancers are defined by enrichment in H3K27 acetylation (H3K27ac) and H3K4 monomethylation (H3K4me1) epigenetic marks. Particularly, while primed and active enhancers are both characterised by H3K4me1 enrichment, H3K27ac mark allows to distinguish active enhancers [6-8].

The transcription factors (TFs) required for proper gene transcription are in continuous and dynamic competition with histones, to modulate chromatin accessibility [4]. Finally, DNA methylation is also a crucial mechanism for proper regulation of gene expression, which is predominant in palindromic CpG dinucleotides in mammalian genomes (CpG islands). Active gene promoters are typically unmethylated, correlated with the requirement for this type of sequences to be accessible to TFs. Opposingly, heterochromatin and several inactive gene promoters are enriched in CpG methylation [4].

a) Transcriptional regulation and cis-regulatory elements

The genome contains genes that are transcribed to copies of ribonucleic acid (RNA) molecules, which are then translated into proteins – functional macromolecules that act on different cellular processes [2]. Transcription is thus one of the first steps on which different factors can act and regulate gene expression.

In eukaryotes, gene transcription starts in the cell nucleus upon binding of RNA Polymerase II (Pol II) to the gene promoter region – canonically defined by a conserved sequence that composes the TATA box. The functional recruitment of Pol II is ensured by the stepwise assembly of several co-factors, in turn stabilized by binding of TFs and chromatin modulators [6]. When the transcription initiation complex is formed, Pol II starts synthesizing a molecule of messenger RNA (mRNA), by adding nucleotides complementary to the DNA template. The newly synthesized mRNA is then transferred to the cell cytoplasm, where it is translated into an amino acidic sequence establishing functional proteins [2, 9].

The human genome is constituted not only by coding DNA used as template for synthesis of proteins, which represent 2% of the genome, but also by non-coding DNA. This non-coding DNA, previously referred as “junk DNA”, composes 98% of the entire human genome. More recently it has been shown that the non-coding DNA comprises several regulatory sequences required for transcriptional regulation [9]. Moreover, almost half of the non-coding human DNA is transcribed to several types of RNA molecules, such as transfer and ribosomal RNAs (tRNAs and rRNAs, respectively) that are part of the machinery involved in regulation of gene expression [2, 9].

Although the mechanism of Pol II binding to the promoter region of a gene is fundamental to initiate transcription, the regulatory mechanisms that act at the transcriptional level are even more complex and dependent on several factors. The complexity associated to specific gene expression patterns, both in space and time, is dependent on the congregated action of all the regulatory elements that compose a gene’s regulatory landscape [10]. The regulatory landscape of a human gene usually includes several cis-regulatory elements (CREs). CREs are non-coding DNA sequences characterised by transcription factor-binding sites (TFBS), which are able to physically interact with gene promoters by chromatin loops, regulating gene expression in a specific manner. These elements can be classified in two major classes, according to their position regarding the site where gene transcription is initiated. The proximal CREs consist mostly on gene promoters, while the distal CREs include distal regulatory sequences, such as enhancers and insulators [9].

(1) Enhancers

Enhancers are described as genomic regions that are able to interact with gene promoters, increasing gene transcription [11]. The first reference to enhancers emerged from molecular cloning studies on which DNA sequences from the SV40 virus were demonstrated to increase the transcription levels of the rabbit haemoglobin *β-globin* gene cloned in the same vector [12]. Enhancers are regulatory elements acting in *cis*, through mechanisms that are independent on the orientation of its sequence. Moreover, the function of enhancers also seems to be independent of the distance between them and the promoter of their target genes, as it has been illustrated in the case of the vertebrate ZRS enhancer controlling the *Sonic hedgehog* (*Shh*) gene located one megabase (Mb) away [13].

Enhancers are typically mapped in intergenic sequences, but they can also be located in introns or exons of genes, either their own target gene or genes of the neighbourhood. These regulatory elements respond to intrinsic and external stimuli [6, 9]. The function of an enhancer depends on the binding of specific TFs, capable to recruit additional transcriptional factors and chromatin remodelers, which enable the physical interaction between enhancer and promoter. For instance, Mediator constitutes an evolutionary conserved protein complex recruited by TFs bound to enhancers, that is able to directly interact with Pol II and other transcriptional co-factors, being a transcriptional key-regulator [14]. The interaction between proteins bound to the enhancer and the ones bound to the promoter seems to dictate the activation of transcription of the target gene [11]. Thus, enhancers are regulatory elements that define precise spaciotemporal gene expression, particularly crucial in vertebrate development [9].

(2) Insulators

Insulators are CREs characterised by an ability to protect genes from transcriptional activation from enhancers belonging to other regulatory landscapes, also known as “enhancer blocking effect” [15]. This way, insulators might define genomic regions in which sequences are more likely to communicate with each other, while insulated from other adjacent sequences, functioning as regulatory boundaries [16]. Insulators characterised as boundary elements usually harbour binding sites of insulator proteins. In vertebrates, the key protein binding to insulator sequences is the CCCTC-binding (CTCF) protein [11]. CTCF is a DNA-binding protein containing a domain formed by 11 central zinc finger proteins. Moreover, insulators bound by CTCF proteins are highly accessible yet they are sequences contained within regions characterised by a high occupancy of nucleosomes [15, 16]. The DNA sequence at the 5′ end of the chicken β-globin *locus* (5’HS4), is a 1,2-kb DNA sequence upstream the β-globin *locus* and constitutes one of the first sequences described as an insulator [17].

Data from chromatin conformation assays demonstrated that the genome is organized in topological associating domains (TADs) [11]. TADs are defined as genomic regions characterised by long-range interactions between promoters and distal enhancers [18]. This type of genomic arrangement is characterised by preferential interactions between *loci* of the same TAD, defined by insulators that block interactions from *loci* that locates outside that TAD. Thus, the 3D organization of the genome

modulates gene expression [10, 11]. TADs boundaries are defined by the presence of CTCF and cohesin binding regions. CTCF complexes bind to specific sites, recruiting cohesin. The classical role of cohesin is associated to sister chromatids cohesion during mitosis and DNA repair by recombination [19]. More recently, cohesin has been proposed to act downstream the CTCF complex on chromatin modulation network, allowing chromatin to fold into loop structures [4, 20]. Different models have been proposed regarding the mechanisms of TAD formation, being the loop extrusion model the common consensus for vertebrates (**Figure 1**) [21]. In this conformational model, the dynamic binding of TFs results in physical proximity between regulatory genomic regions, conformations held by the cohesin complexes. Cohesin is recruited to the DNA by CTCF, beginning an extrusion DNA loop that extends until cohesin detects an adjacent occupied CTCF-binding site [20]. Meanwhile, Polymerase II binds to gene promoters and transcription can be activated by proximal enhancers [11]. This model adjusts well in the light of a mechanism of promoter competition, where the proximity between promoter and enhancer acts as an advantage to mediate interactions [11, 22]. Regarding high order structure, the topological organization defined by TADs helps to explain the favouring of some long-range interactions. The presence of TADs boundaries defines regions where genes and CREs contained within seem to preferentially interact with each other (**Figure 1**), therefore establishing genomic regulatory regions [22]. Enhancers appear to play a limited role regarding specificity for their target promoters, whereas the establishment of TADs acts a determinant factor in the formation of specific interactions of those enhancers with target promoters [19]. The dynamic binding of TFs to enhancers, together with the distribution of CREs in TADs are suggested to regulate specific promoter-enhancer interactions required during development and in different tissues [21, 22].

b) Characterization of the chromatin

Several techniques have been developed to study different properties of the chromatin, from analysis of specific *loci* to the “omics” included in genome wide studies. Next-generation sequencing (NGS) enabling to process and organize high-throughput data from genomic, transcriptomic and epigenomic techniques underwent a fast development during the early 2000’s. Bioinformatics analysis of genome-wide data greatly contributed to understand the molecular basis of complex human diseases [23].

- **DNA accessibility**

In order to enable the binding of TFs to regulatory sequences, those sequences are required to be accessible, meaning that they usually need to be contained within nucleosome-free regions [5]. Several techniques have been developed to identify accessible or open regions of chromatin. Chromatin digestion by micrococcal nuclease (MNase) was one of the first technique employed to investigate chromatin accessibility. This nuclease is able to cut double-stranded DNA linking sequences between nucleosomes, providing information on nucleosome location [24]. Higher resolution can be obtained by other technique that allows identifying accessible chromatin regions – *DNase* I-Seq. This allows to identify genomic regions sensible to the activity of the hypersensible endonuclease *DNase* I, followed by sequencing of the digested fragments [25]. An alternative technique used to this end is the Formaldehyde-Assisted Isolation of Regulatory Elements coupled to high-throughput sequencing (FAIRE-Seq). FAIRE enables to identify open chromatin regions, through the isolation of nucleosome-free DNA from the occupied DNA and assessment of actively transcribed sequences [26].

The mapping of accessible chromatin regions through the techniques described above requires high input of DNA samples, which is not always possible. A technique requiring lower amount of chromatin is the assay for Transposase Accessible Chromatin, followed by high-throughput sequencing (ATAC-Seq). This type of assay makes use of a highly active transposase (Tn5) to evaluate the accessibility of certain genomic regions. The transposase binds to accessible sites and fragments DNA, then enabling the insertion of sequencing adapters [18]. Sequencing results are then analysed to create maps of open chromatin, reflecting nucleosome position and accessibility [27].

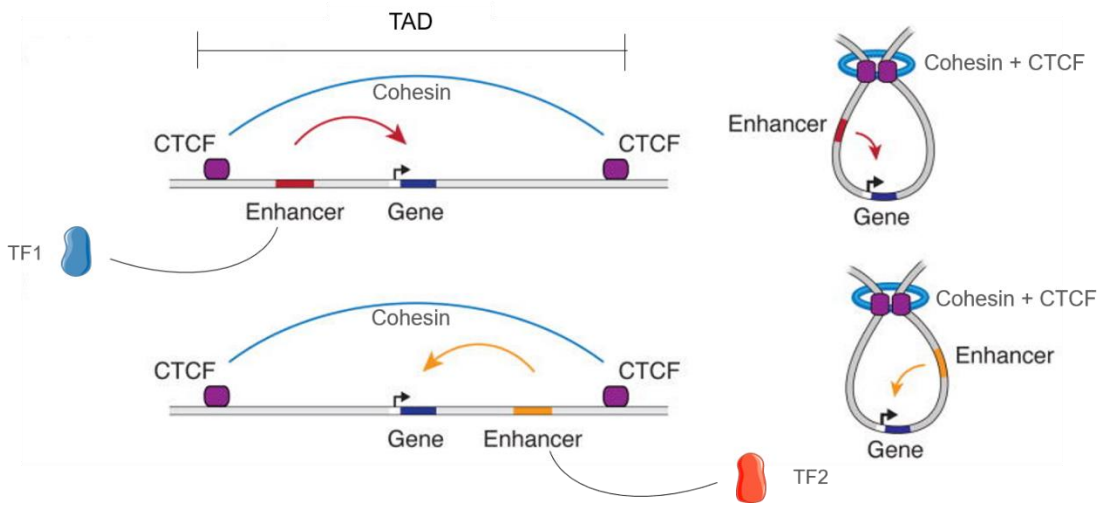

Figure 1. Topologically associating domains (TADs). TADs define genomic regions where promoters and enhancers interact. The binding of specific TFs explains helps to explain the variety of interactions that could be detected within a TAD. Adapted from Hnisz, D et al. 2016.

- **TF binding**

The activity of transcriptional regulatory sequences is also in part modulated by interactions between DNA and DNA-binding proteins [5]. In order to identify *in vivo* binding of a given TF to a DNA sequence, Chromatin Immunoprecipitation coupled to high-throughput sequencing (ChIP-Seq) is the prime technique. In this assay, chromatin is sonicated and immunoprecipitated with specific antibodies, followed by massive parallel sequencing. ChIP can be used to detect the biding of TFs and also histone modifications; this technique requires previous information regarding potential transcription factors that might bind to the regulatory sequences in test, so that specific antibodies can be selected [28].

- **DNA conformation**

Transcriptional cis-regulation by long-range interactions, depends greatly on the 3D structure of the chromatin [29]. This structure can be evaluated by Chromatin Conformation Capture (3C) and associated technologies. Reported for the first time in 2002, the 3C technique relies on formaldehyde crosslinking of chromatin, followed by digestion with restriction enzymes and intramolecular ligation of cross-linked fragments [30]. The ligated fragments are hybrid DNA molecules, which reflects the frequency of the *in vivo* interactions mediated by protein complexes. Development of this technique allowed using the ligated DNA as template in quantitative PCR reactions to identify the physical distance between desired *loci* in the 3D structure of the chromatin in the cell's nucleus. This way, 3C constitutes a quantitative technique to detect the frequency if interactions among 2 *loci* [31].

Based on 3C, several technologies have been developed and coupled with Next Generation Sequencing. Most of these technologies present common initial steps, comprising crosslinking, digestion and ligation of digested chromatin fragments [29]. The Circularized Chromosome Conformation Capture (4C) technique offers several advantages in comparison to 3C. The establishment of circular fragments enables to capture conformations that reflect genome-wide interactions between unknown DNA sequences and the viewpoint [32]. Other important breakthrough in chromatin conformation analysis was the establishment of the Hi-C protocol. Using an all-to-all approach, Hi-C differs allows identifying simultaneously cis- and trans-interactions in an unbiased way, that is, without a specific and pre-selected viewpoint [33]. For this reason, Hi-C is considered an unbiased genome-wide assay. The Hi-C technique could also be applied to more specific purposes, such as the promoter capture Hi-C (pcHi-C). pcHi-C

enables to identify interactions between targeted promoters and CREs with high-confidence, allowing to create high-resolution chromatin interactome maps [18].

c) The contribution of chromatin organization and cis-regulation to disease development

Complex or multifactorial diseases, such as cancer and diabetes, result from a combination of genetic and environmental factors. Moreover, complex diseases are defined by genetic traits derived from different *loci* that characterise disease predisposition. Human complex diseases are frequently associated to the dysregulation of gene transcription. The concomitant lack of point mutations in the coding sequence of given genes highlights the relevance of genetic variants in non-coding sequences to the development of those diseases [10, 34]. Indeed, the overlap between non-coding polymorphisms associated to disease, identified by genome wide association studies (GWAS), and identified regulatory elements, such as enhancers, suggests an impact on transcriptional regulation [34-36]. In spite of the corroborated importance of cis-regulation to the development of human diseases [37, 38], a formal demonstration and the mechanisms associated to the impairment of normal cis-regulatory networks remain elusive.

Transcription regulation mechanisms correlate with an increased complexity in gene expression patterns, being the source of differential gene expression throughout vertebrate development [2]. Consequentially, several diseases arise from the dysregulation of regulatory mechanisms that interfere with the complex synchronization of cellular functions [6, 10, 11]. Therefore, it is essential to describe human chromatin conformation maps in order to study the molecular mechanisms inherent to the pathogenesis of those diseases.

The evolution of chromatin conformation assays, complemented with the functional analysis of relevant *loci* associated to complex diseases, brings new insights into the regulatory mechanisms underlying pathogenic pathways. A noteworthy example related to the main focus of this thesis – the study of the pancreas – came in 2017, with the online release of the Islet Regulome Browser [23]. This tool gathers genomic, epigenomic and transcriptomic results from genome-wide studies from human islets, predicting the genomic location of CREs active in human islets: in detail, this browser integrates data from FAIRE-Seq, ChIP-Seq, mRNA-Seq, ATAC-Seq and pcHi-C; moreover, it includes data from human adult pancreatic islets as well as pancreatic progenitors cell types [7, 18, 23, 28, 39]. This platform allows the integration of processed

data from different sources in a fast and easily way [23]. In summary, the Islet Regulome Browser enables the scrutiny of regulatory networks and the selection of *loci* for experimental designs, namely in diabetes research [23].

2) Pancreas

a) Pancreas organization

The human pancreas is a glandular organ with important roles on both digestive and endocrine systems [40]. Functionally, it can be divided in exocrine and endocrine compartments.

The exocrine pancreas is composed by epithelial acinar cells, responsible for the production of digestive enzymes – trypsin, chymotrypsin, amylase and lipase – and ductal cells, organized into ducts that converge to the main pancreatic duct. This structure enables the secretion of such enzymes to the duodenum [38, 42]. In cases of dysfunction of the acinar cells or obstruction of the pancreatic duct, the secretion of digestive enzymes can get compromised, as well as the digestive process. This type of dysfunctions can lead to the development of pancreatic inflammation and tumorigenesis, namely adenocarcinoma [41, 42].

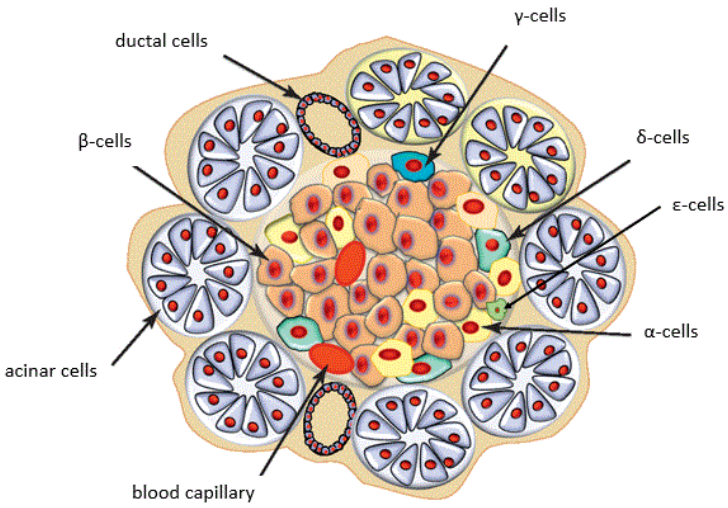

Figure 2. Schematic representation of islets and cell types present in human pancreas. Adapted from Efrat, S., & Russ, H. A. (2012).

The endocrine pancreas is responsible for the production of several hormones involved in glucose homeostasis [43], which are released to the blood vessels [44]. The

endocrine cells are organised in clusters imbedded in exocrine cells (**Figure 2**). These clusters are called islets of Langerhans and comprise α-, β-, δ-, ε- and γ-cells that produce glucagon, insulin, somatostatin, ghrelin and pancreatic polypeptide, respectively [45]. Glucose homeostasis is primarily ensured through a correct balance between glucagon and insulin. Proper production of these enzymes is regulated by a feedback loop to maintain this balance. While insulin decreases blood glucose levels, glucagon exerts the opposite effect [2]. Moreover, somatostatin inhibits the production of both insulin and glucagon. Impairment of normal endocrine cell function results in abnormal levels of pancreatic hormones, and therefore in dysregulation of glucose blood levels – a pathogenic phenotype that characterises diabetes mellitus [45].

b) Vertebrate pancreatic development

The vertebrate pancreas is derived from the foregut of the embryo, which undergoes complex signalling pathways to induce the formation of the ventral and dorsal pancreatic buds [45]. Most part of the information gathered about the induction and development of the vertebrate pancreas regards genetic studies in mice, which are also the basis of the regulatory networks that will be explained hereafter.

The pancreatic buds are composed by proto-differentiated Multipotent Progenitor Cells (MPCs). Right in the first stages of pancreatic development, MPCs express the *pancreatic and duodenal homeobox 1* (*Pdx1*) and the *pancreas transcription factor 1A* (*Ptf1a*). The relevance of Pdx1 and Ptf1a is highlighted by homozygous mutations of each gene in mice, which induce pancreatic agenesis [42]. Pdx1 and Ptf1a are suggested to be crucial TFs required to induce proper development of the pancreas, with *Pdx1* even being earlier expressed in the primitive gut [46].

Other TFs have been reported to act upstream *Pdx1* and *Ptf1a*. The presence of binding sites for SRY-Box9 (Sox9) on the promoter of *Pdx1*, as well as its regulatory function regarding the expression of *Hepatocyte nuclear factor 1 b* (*Hnf1b)*, *Hepatocyte nuclear factor* (*Hnf6)* and *Forkhead box A2* (Foxa2), suggests that Sox9 plays an important role in the pancreatic regulatory network. Moreover, later during development Sox9 inducts the progression of progenitors towards the exocrine fate [47].

The establishment of the pancreatic buds is followed by pancreatic morphogenesis, a tightly coordinated process that involves cell polarization and epithelial stratification and arrangement into microlumen structures [46]. In rodents, pancreatic organogenesis is characterised by two consecutive temporal transitions. The first one comprehends

pancreatic induction, budding and fusion, along with the formation and proliferation of a pool of MPCs and development of the microlumen. This stage is characterised a first wave of endocrine cells formation, during which the first α- and β-cells are produced [43]. Upon pancreatic budding, multipotent MPCs located in the "tip" domain of the developing pancreas, are characterised by expression of *Ptf1a,* while cells located in the "trunk" domain express homeobox protein NK6 homeobox 1 (Nkx6.1). Nkx6.1^{+} cells are bipotent, as they are committed whether to ductal or endocrine cells [45]. During the second transition, the microlumen undergoes a morphogenic remodelling process to constitute the luminal network and the second wave of endocrine cell formation, the progenitor cells.

Throughout pancreas development, the balance between levels of the bHLH TF neurogenin 3 (Neurog3) and the Notch signalling factors seem to be determinant to cell differentiation. During the bud stages, mice pancreatic progenitors transiently express *Neurog3* for a time period of 12 to 24h. *Neurog3* expression undergoes two transient periods, correspondent to the two waves of pancreatic organogenesis [48]. The expression of this gene is upregulated by Notch signals , as loss-of-function studies in mice mutants showed that high levels of *Notch* induce the expression of *Sox9* and, consequently, the expression of *Neurog3* [49]. Notch is suggested to be a crucial orchestrator of pancreas development, being required to the proliferation of MPCs and the maintenance of the progenitor state of these cells. Moreover, Notch activity in progenitors regulates the patterning of the pancreas in "tip" and "trunk" domains, by activating Nkx6.1 expression and repressing Ptf1a [45].

Furthermore, the Notch pathway is involved in the determination of cell lineage commitment. Neurog3^{high} cells seem to be deviated from the progenitor state, by induction of cell cycle exit, and determined to an exocrine fate – in these cells, the high levels of Sox9 induce the differentiation of ductal cells, while the maintenance of high levels of Ptf1a leads cells towards an acinar fate. On the contrary, low levels of Notch induce not only the expression of *Sox9*, but also the expression of *Hes1*, that act together on the downregulation of *Neurog3*. Neurog3^{low} cells transiently re-enter cell cycle before the second transition, giving rise to unipotent endocrine precursors [49, 50]. Moreover, Neurog3 undergoes posttranslational regulation from the Notch signalling pathway, through protein destabilization [50]. The congregated action of endocrine-associated TFs, along with the downregulation of progenitor-associated markers, is then suggested as a determining factor to proper endocrine differentiation and function [43, 48].

Additionally, the environmental niches in which the progenitors are contained also influence the fate of progenitor cells. For instance, after embryonic day (E) 12,5 of mouse development, the epithelial stratification induces the restriction of MPCs to the acinar fate and the clustering of endocrine progenitors that will form the islets of Langerhans [43, 48].

- **Human pancreatic development**

The human pancreas arises from the ventral and dorsal buds of the embryo foregut, which later fuse in order to constitute a unique organ. Human pancreatic development presents several common aspects with the murine one [44]. During early embryogenesis, transient expression of *NEUROG3* is present in humans, as it is in rodents. The crosstalk among NEUROG3 and the Notch signalling is also implied in the maintenance of pancreatic progenitors and cell differentiation [43]. In contrast, a single wave of endocrine cells differentiation occurs and the pancreatic morphology acquired upon cell rearrangement also differs [43]. MPCs begin to proliferate already before evagination of the pancreatic buds, establishing a pool of progenitors throughout embryogenesis [51]. Endocrine or exocrine cell commitment of human pancreatic cells depends on the coordinated action of a TF network, including NEUROG3, FOXA2, PDX1, PTF1A and NKX6-1 [38, 52]. Remarkably, *PDX1* is described to be broadly expressed in pancreatic progenitors and to continue to be expressed during adulthood [38].

c) Zebrafish as a model organism to study pancreatic development and function

- **Zebrafish as a model organism**

The zebrafish (*Danio rerio*) presents multiple advantages as a vertebrate model organism, when in comparison to mice. It is easy and low-cost to maintain. It has high fertility rates. The external fertilization facilitates genetic manipulation and transgenesis in embryos, through microinjection [51, 52]. Its development is fast and embryos are transparent, allowing the track of fluorescent proteins and organogenesis processes *in vivo*. Additionally, zebrafish shows a short generation time, where fish reach sexual maturity within 3 months, which enables to follow heritability and to create stable transgenic lines in short timeframes [53]. Finally, the zebrafish genome includes more than 26 thousand annotated genes, from which 69% are human orthologous [40]. This is accompanied by the conservation of most of developmental processes and gene regulatory networks between the two species [52].

Zebrafish transgenesis is very suitable to the study of transcriptional regulation. A notable method consists on the validation and characterisation of CREs through cloning in shuttle vectors and microinjection of zebrafish embryos [10, 15]. In the case of enhancers, the sequence of interest is cloned upstream of a minimal promoter, that alone is not able to trigger transcription of an *in vivo* reporter gene located downstream, such as the one encoding a green florescent protein (GFP; [15]). This vector is then integrated in the zebrafish genome and if the reporter gene is observed to be expressed, usually by direct *in vivo* microscopic observation, the cloned fragment is validated as an enhancer. Other types of CREs can also be easily tested using transgenesis in zebrafish, as insulators. In this case, putative insulators are cloned in between an enhancer and a promoter located upstream an *in vivo* reporter gene. Upon the introduction of this vector in the zebrafish genome, if the activity of the enhancer is impaired, but not affecting the activity of the promoter, the cloned fragment is validated as an insulator [15].

Further development of transgenesis methods allowed to increase the size of potential regulatory genomic fragments. While transgenesis based on small plasmids is performed through isolation of the genomic sequence of interest and assessment of its regulatory function on the expression of a reporter gene, artificial-chromosome type plasmids, such as Bacterial Artificial Chromosomes (BACs), show higher capacity for cloned genomic fragments [54]. Integration of larger genomic fragments in model organisms, namely zebrafish, is possible due to constitution of transposable elements. The most widely used system of transgenesis is based on the cut-and-paste activity of the *Tol2* transposon [55], leading to high efficient random integrations in the zebrafish genome, upon microinjection of one-cell stage zebrafish embryos [56]. BAC transgenesis allows to overcome several limitations of transgenesis with small plasmids: first, it allows to study regulatory sequences without disturbing their original genomic context, highlighting complex gene regulatory mechanisms; moreover, large transgenes are associated to more consistent expression levels, showing less unspecific effects due to interaction of the transgene with surrounding genomic elements in the host genome (referred as “position effect”) [54, 56].

Because many of the zebrafish genes are conserved in humans, loss-of-function assays have been fundamental to address their function. Mutagenesis, and other types of loss-of-function assays of zebrafish genes have been performed in many different genetic studies, being a foundation for both forward and reverse genetics. As a relevant example to this thesis, O’Hare and colleagues made use of the zebrafish model to identify candidate genes that could be directly implicated with T2D. This was performed

in zebrafish by the loss-of-function of several candidate genes, known to be located in human T2D-associated *loci,* and analysing its impact in β-cell mass [57]. Thus, zebrafish is an extremely reliable model organism to the research of diabetes and to the development of new therapies.

- **The zebrafish pancreas**

The zebrafish pancreas presents several common aspects with the pancreas of mammals. Phylogenetic studies have shown that the digestive and endocrine systems have fused to form the pancreas throughout fish evolution [40]. Zebrafish shares not only a conserved pancreatic organization and cellular composition, but also similar developmental processes regarding the pancreatic development of mammals [58].

The mature zebrafish pancreas reassembles the human organ, consisting on islets of endocrine cells dispersed through the acinar cell mass of the exocrine portion [41, 51]. In both organisms, the main functions of the endocrine and exocrine pancreas are conserved consisting on production of hormones regulating glucose homeostasis and digestive enzymes released to the pancreatic ducts, respectively [53].

The zebrafish pancreas derives from the ventral and dorsal buds, as it happens in mammals [53]. The pancreatic buds arise from a *Pdx1*-expressing cell domain in the foregut endoderm. The patterning of the zebrafish embryo foregut correlates with the gradient of *Pdx1* expression, which together with BMP signalling contributes to determine the liver and the pancreatic progenitors. The most exterior 2-cell layer of the foregut gives rise to the liver, while the intermediate and most interior layers contribute to the exocrine and the endocrine pancreas, respectively [59]. Moreover, the Notch signalling pathway implied in pancreatic development of mammals is also conserved in zebrafish. Studies performed in a Notch reporter line proved that high levels of this factor correlate with quiescence, while low levels primarily lead to the proliferation of progenitors, and consequently to cell cycle exit and endocrine differentiation when in continuous downregulation [60]. *Hnf1ba* is also reported to be associated to zebrafish pancreatic development, along with the Wnt signaling pathway. In divergence to mice, the zebrafish mutant allele *hnf1ba*^{s430} is known to recapitulate the phenotype of MODY5, a form of diabetes arisen from the mutation of the human gene [61].

Despite the similarities among pancreatic vertebrate development, each organism presents particular characteristics. For instance, the cells composing the dorsal bud of the zebrafish pancreas are not multipotent cells as in mammals; instead, the lineage commitment of the first endocrine progenitor is already restricted to the principal islet

[53]. In contrast to mice, the zebrafish endocrine progenitors do not express *Neurog3*; instead, Ascl1b and NeuroD1, two distinct TFs belonging to the same bHLH family, have been shown to replace Neurog3 functions. The importance of the cooperative action of Ascl1b and NeuroD1 is highlighted by the complete absence of the endocrine pancreas upon double knockdown by morpholino [62]. *Ascl1b* is expressed in both pancreatic buds at 10 hpf, where it induces the appearing of the first endocrine progenitors expressing *Sox4b*. Cells expressing both *Ascl1b* and *Sox4b* undergo upregulation of *NeuroD1* [58]. In endocrine progenitors, high and low levels of NeuroD1 determine the differentiation of α- and β-cells, correspondingly [63].

d) Pancreatic diseases

The aetiology of different human pancreatic diseases, is associated to variable predisposition traits, from genetic to environmental factors. This is the case of chronic pancreatitis, diabetes and pancreatic cancer [41]. Moreover, these complex diseases arise from a combined effect of mutations in different *loci*, which attribute a genetically heterogeneous background of susceptibility [64].

Pancreatic diseases have been broadly associated with mutations on non-coding regions of the genome [10, 34]. In particular, non-coding mutations on cis-regulatory elements of key pancreatic genes seem to impair the correct gene regulatory networks required for normal pancreatic development [7]. Furthermore, some cis-regulatory mutations were shown to lead to total absence of the human pancreas [34-36].

The mechanisms regulating vertebrate pancreas development have been studied in the last decades, namely the gene regulatory networks that were discussed above. Although extensively investigated, the molecular and cellular processes implied in pancreas morphogenesis and organogenesis remain elusive. For instance, cell fate determination and cell polarization during epithelial remodelling and islet has not been fully characterised, and is fundamental to understand their impact on pancreatic diseases. Further knowledge on gene cis-regulatory mechanisms required for proper pancreatic development might help highlighting the impact of cis-regulatory mutations to the onset of pancreatic diseases, as well identifying novel therapeutic targets of human pathologies, from pancreatitis to diabetes and pancreatic cancer [45, 48].

3) Diabetes mellitus

Diabetes is a multifactorial disease, being a reflection of the combined effect of genetic predisposition, environmental factors, aging and lifestyle [65-67]. The disease presents worldwide incidence, being reported to affect more than 422 million adults in 2016, according to the Global Report On Diabetes from World Health Organization (WHO) [68]. When undiagnosed or defectively managed, diabetes can also be associated to vascular difficulties, such as cardiovascular disease [67, 68].

Diabetes mellitus includes a range of metabolic disorders characterised by hyperglycaemia [66, 69]. Hyperglycaemic traits arise from dysregulation of glucose homeostasis, which depends essentially on the action of insulin and glucagon through feedback mechanisms [1].

Of note, diabetes seems to be considered as a risk factor for pancreatic adenocarcinoma and recent epidemiological studies suggest that the causality also works in the reverse direction [41]. This correlation highlights not only the relevance of a coordinated function between exocrine and endocrine pancreas, but also the need to unravel the molecular mechanisms triggering pancreatic dysfunction.

The first evidence of a correlation among high blood sugar levels and pancreatic malfunction was reported in 1889, by Joseph von Mering and Oskar Minkowski [70]. A pioneer medical and biochemical investigation regarding pancreatic extracts led to the finding of the molecule responsible for glucose homeostatic control [71]. The discovery of insulin, in 1921, was attributed to Frederick Banting and Charles Best, a finding awarded with the Nobel Prize in Physiology, in 1923 [70]. Further studies started to focus on the molecular mechanisms underlying insulin production. Diabetes is a world epidemy and the disease complexity has been continuously highlighted, with the attribution of genetic and environmental susceptibility factors [72].

a) Different forms of diabetes

Diabetes mellitus is divided in four main forms, differing on the predisposition risk factors, age of onset and aetiology [67]. Based on these criteria, diabetes mellitus comprises type 1 and type 2 diabetes, gestational diabetes and other specific forms of diabetes, including disorders arising from both genetic and non-genetic molecular origins [67, 73].

Regarding the last category, first reports on monogenic forms of diabetes are dated to the 90s, with the first description of Maturity-onset diabetes of the young (MODY) [74]. MODY is a rare autosomal dominant form of diabetes, which is characterised by an onset before the age of 25 years [75, 76]. Currently, there are 13 genes associated with this disease onset (**Table 1**), most of them associated to the regulation of insulin secretion as a response to glucose bloodstream levels [76]. Consequently, mutations in these genes lead to decreased insulin secretion by β-cells [75, 76]. Recent studies have been suggesting that predisposition to MODY is increased by other genetic, as well as environmental factors [76, 77]. The implication of different genetic variants in MODY pathogenesis, could be used as a tool to the study of multifactorial types of diabetes [76].

Table 1. **List of MODY types and associated genes**. Data from OMIM – Online Mendelian Inheritance in Man [78].

Type	MODY-associated genes	Specific phenotypic traits
MODY1	Hepatocyte nuclear factor-4-alpha gene (HNF4A)	
MODY2	Glucokinase gene (GCK)	
MODY3	Hepatocyte nuclear factor-1alpha gene (HNF1A)	
MODY4	Pancreas/duodenum homeobox protein-1 gene (PDX1)	
MODY5	Hepatic transcription factor-2 (TCF2)	Atrophy of the pancreas and forms of renal disease
MODY6	Neurogenic differentiation 1 (NEUROD1)	
MODY7	Kruppel-like factor 11 (KLF11)	
MODY8	Carboxyl-ester lipase (CEL)	Exocrine dysfunction
MODY9	Paired box gene 4 (PAX4)	
MODY10	Insulin (*INS*)	
MODY11	Tyrosine kinase gene (BLK)	

MODY13	Potassium voltage-gated channel subfamily J member 11 (KCNJ11)
MODY14	Adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 1 (APPL1)

Aside from the monogenic forms, the most represented types of diabetes are polygenic, namely type 1 and type 2 diabetes (T2D) [74, 79]. Type 1 diabetes is characterised by pancreatic β-cells destruction due to autoimmune dysfunction. Affected patients display life-long dependence on insulin administration [67]. Instead, T2D is caused by the impairment of normal pancreatic function due to deficient insulin production by β-cells or by insulin resistance [57, 66, 69].

b) Type 2 diabetes (T2D)

Type 2 diabetes (T2D) is thought to affect around 90% of diabetic patients. This number seems to be associated to the nature of its exclusion diagnosis – T2D diagnosis outcomes from cases of hyperglycaemic patients showing no evidence of autoimmune deficiency (distinctive of type 1 diabetes) and no evidence for monogenic forms of diabetes [52, 65, 80]. Despite of the high prevalence among diabetic patients, the disease complexity makes T2D frequently undiagnosed. Moreover, the molecular mechanisms behind T2D are still poorly understood [66, 68].

The complex genetic nature of T2D has been confirmed by several Genome-Wide Association Studies (GWAS) [65, 66, 80]. GWAS identify single-nucleotide polymorphisms (SNPs) linked to phenotypic traits [81]. This type of genome-wide analysis allows to understand the combined contribution of several genomic *loci* to diabetes development, along with its association to both coding and non-coding sequences [57, 69, 81]. Interestingly, the majority of T2D-associated SNPs are mapped to non-coding sequences of the human genome, suggesting that the impact on disease development depends on the disruption of epigenetic mechanisms controling gene transcription [37]. Recent studies have detailed that these SNPs are enriched in chromatin regions that display marks of cis-regulatory elements (CREs), some of which showing activity specifically in pancreatic islets [82]. Moreover, the majority of these SNP containing non-coding sequences were shown not to be in linkage desiquilibrium with coding sequences, meaning that there is no evidence of preferential segregation of those non-coding sequences in combination with coding genes [36]. Alltoghether, these

colected data indicated that the role played by these SNPs on a disease-associated phenotype might be related to cis-regulatory mechanisms [34].

Another layer of investigation relates to expression Quantitative Trait *Loci* (eQTLs) [83]. These *loci* can be defined as genomic regions containing polymorphisms and described to influence gene expression levels [34]. Analysis of eQTLs helps to gain insights on how sequence variation is linked to changes in gene expression, and consequently, having the potential to produce disease related phenotypes [34, 83]. Complementing each other, the analysis of eQTLs allows to recognise causal SNPs identified in GWAS, that can have functional relevance, especially in the context of pathogenic phenotypes [34, 84].

4) Pancreatic and duodenal homeobox 1 (PDX1)

Pancreatic and duodenal homeobox 1 (*PDX1*), also known as *insulin promoter factor-1* (*IPF1*), is a master regulator of pancreatic development and function, which is conserved through vertebrates [85]. The human *PDX1* gene is composed by an 852 base pairs (bps) coding sequence, constituting 2 exons, and it is localized in chromosome 13 of the human genome.

The sequence of *PDX1* encodes a 283 amino acids (aa) protein. PDX1 protein is a TF responsible for the transcriptional activation of several genes, including *insulin*, *somatostatin, glucokinase*, *islet amyloid polypeptide* (*IAPP*) and *glucose transporter type 2* [52]. This protein includes a N-terminal transactivation domain (13-73 aa), followed by anti-type hexapeptide (118-123 aa), which is required for the interaction with PBX1. The C-terminal region of the protein contains a homeobox domain (149-203 aa), including 3 helices and a nuclear localization signal (NLS, 197-203 aa). The homeobox on PDX1 protein is required for DNA-binding, but it also functions as mediator for protein-protein interactions between PDX1 and its interactors [86, 87].

The human PDX1 protein is conserved in vertebrates, including mice and zebrafish. Mice Pdx1 shows 80% sequence identity with the human protein, whereas zebrafish pdx1 shows 54% of identity (BLAST analysis, [88]). In particular, the homeobox domain, a central component regulator of developmental mechanisms throughout eukaryotes, is highly conserved among PDX1 orthologous proteins. Notably, the homeodomain of the zebrafish pdx1 protein is 95% identical to the one contained within the mammalian orthologs [46, 89].

The role of PDX1 in transcriptional regulation relies on the specific targeting of genes to be expressed. The mechanism through which PDX1 binds to target genes depends on the protein homeodomain – that enables the recognition of A/T-rich sequences. One of its direct target genes is *insulin*, which transcription is enhanced by PDX1 in mature islets. In contrast, the occupancy of the *insulin* promoter in ductal cells by a nucleosome seems to be responsible for the inaccessibility of the insulin promoter to PDX1 in this cell type [90, 91].

a) PDX1 on pancreatic development and adult function

Among different TFs, PDX1 is known to control the initial steps of pancreas development and differentiation, as well as maintaining the function of differentiated pancreatic cells [53]. Human *PDX1* has been described as broadly expressed in pancreatic progenitors around 4 weeks post-fertilization and to continues to be expressed during adulthood [38]. Following chromatin remodelling driven by pioneer factors that act in human embryonic development [92], as FOXA1 and FOXA2, to activate several pancreatic enhancers, PDX1 specifically binds to those to determine cell lineage fate [52, 85, 93]. Acting downstream of the Notch signalling pathway on progenitor cells, PDX1 is required to the differentiation of the endocrine pancreas. Upon the differentiation of endocrine cells, *PDX1* is expressed in mature β-cells. This expression is crucial to maintain the insulin levels produced by these endocrine cells, and consequently, to ensure a proper regulation of glucose homeostasis [93, 94].

The relevance of PDX1 for pancreas development and adult function is conserved in mammals, as well as in other vertebrates including zebrafish [48].

b) PDX1 association to disease

The relevance of PDX1 for pancreatic development, as well for proper mature β-cell function, is demonstrated by the phenotypes arising from different mutations in the *locus*. In humans, homozygous mutations of the *PDX1* gene totally impair pancreatic development, leading to pancreatic agenesis. In case of heterozygous loss of the gene, the low levels of PDX1 are associated to MODY type 4 (MODY4), a phenotype also associated to frameshift mutations on the gene's coding sequence. MODY4 is a form of diabetes characterised by deficient insulin secretion and dysregulation of glucose homeostasis (**Table 1**). More recently, homozygous hypomorphic *PDX1* mutations, leading to decreased gene activity, have been correlated with neonatal diabetes.

Additionally, mutations on PDX1-binding sites located in cis-regulatory elements that control the expression of different *loci* are correlated to T2D susceptibility [87, 92]. Therefore, total loss or partial deletion is sufficient to cause disease, highlighting the importance of PDX1 for proper pancreatic function.

Table 2. **Phenotypes described as result of mutations on the human PDX1 *locus*, as well its mice and zebrafish orthologs.**

	Human	Mice	Zebrafish
PDX1$^{(+/-)}$	Defects in insulin production and glucose homeostasis (MODY4)	Defects in insulin production and glucose homeostasis	Defects in insulin production and glucose homeostasis
PDX1$^{(-/-)}$	Pancreatic agenesis	Pancreatic agenesis	Defects in β-cell function and diabetes

The phenotypes arising from *Pdx1* mutations in vertebrate orthologs resemble the ones described in humans (**Table 2**). In zebrafish and mice, heterozygous *Pdx1* mutants show defects on insulin production and β-cell survival [87]. Homozygous *Pdx1* mutant mice fail to develop a pancreas, resulting in organ agenesis, while zebrafish mutants manage to develop a pancreas, even though they show endocrine dysfunction and diabetic traits [87, 93].

c) Regulation of PDX1

Research on the regulatory elements defining *Pdx1* expression throughout development and in specific cell types, revealed a sequence located approximately 2-kb upstream the transcription start site of the mice *Pdx1* gene that mediates expression in pancreas, stomach and duodenum [95]. This regulatory sequence holds four domains – termed areas I, II, III and IV – which are phylogenetically conserved in mammals. Area III is involved in pancreas-wide expression of *Pdx1* during early bud formation, whereas areas I, II and IV were reported to control restricted gene expression in islets during later development and in the mature pancreas [96]. Transfection assays performed in human β-cell lines showed that areas I, II and IV induce tissue-specific *Pdx1* expression in an independent manner. The transcriptional regulation of *Pdx1* in β-cells is partially mediated by TFs implicated in the development of this cell line, such as FoxA2, MafA,

Hnf1a and Pdx1 itself [95, 97]. Moreover, regulatory areas I–IV are located in mapped regions of open chromatin of β-cells [98].

PDX1 regulation is dependent on upstream and downstream co-factors, as well as chromatin structure. These aspects influence PDX1 levels on distinct cell types, enabling the expression of *PDX1* in early progenitors, followed by specific regulation of gene expression in high levels in insulin-producing cells and in low levels in certain somatostatin-producing and exocrine cells [96].

Another layer of regulation of PDX1 levels is achieved by post-translation modifications, including phosphorylation, sumoylation and glycosylation. For instance, upon the action of kinases on specific residues, phosphorylated forms of PDX1 are targeted for proteosomal degradation. Moreover, observations suggesting a cytoplasmic localization for PDX1 are associated with cytoplasmic sequestration. This regulatory mechanism of PDX1 function responds to physiological conditions, also acting in pathological conditions, such as oxidative stress seen in cases β-cell dysfunction [99-101].

a) Transcriptional regulation by PDX1

The mechanisms through which PDX1 regulates gene transcription have been explored in several studies, the majority of those using mice as model organisms. In these studies, Pdx1 is described to bind to other TFs, including NeuroD1, MafA, Hnf1β and Nk2 homeobox 2 (Nkx2-2), as well to transcriptional co-activators, such as p300 and Bridge 1 [102]. Altogether, the recruitment of co-activators that link Pdx1 to the Pol II transcriptional machinery allows the establishment of protein complexes, which enhance DNA binding affinity through modulation of DNA conformation [103].

The tight transcriptional regulation of *PDX1*, as well as the role of the TF it encodes in several regulatory networks remains elusive. Knowledge of *PDX1* cis-regulatory sequences is essential to understand the complex gene expression patterns of this pancreatic master regulator in specific developmental stages and cell types. Furthermore, the study of *PDX1* transcriptional regulation should allow gaining insight on all the pathways in which this gene is implicated, helping to identify more protein interactors and the mechanisms through which the dysregulation of *PDX1* expression is associated to disease.

5) Hypothesis and main objectives

This work aims to increase our understanding of the impact of human cis-regulatory mutations in the development of T2D-associated traits by studying a human relevant regulatory landscape. Using bioinformatic tools, we selected the *locus* of *PDX1,* that is associated to a monogenic form of diabetes, MODY4, and that contains a non-coding SNP known to be associate to T2D [7]. Our hypothesis is that changes in *PDX1* CREs might lead to its transcriptional disruption causing pancreatic T2D associated phenotypes. To clarify this, we next screened and validated CREs by transgenesis reporter assays, contained in the landscape of PDX1, using the zebrafish as an *in vivo* model. Finally, and since the study of CREs out of their genomic context is very limited to address their contribution to *PDX1* transcription, we aimed to generate a zebrafish transgenic line containing the human PDX1 *locus* with its corresponding regulatory landscape. This line will allow to study the contribution of nucleotide modifications in the regulatory landscape of PDX1 to its transcription. Therefore, this work included four goals to be achieved:

1. Bioinformatic analysis of *loci* of interest: T2D-associated SNPs, prediction of CREs and analysis of gene function;
2. Mapping and *in vivo* validation of CREs within the selected *locus*;
3. BAC recombineering in *E. coli*;
4. Transposon-mediated BAC transgenesis in zebrafish.

Material and Methods

1. *Locus* selection

The selection of a human *locus* of interest was based on publicly available GWAS data; in detail – GWAS datasets from Pasquali, L. et al. (2014), Sun W. et al. (2018), Mahajan, A. et al. (2014), O'Hare, E. A. et al. (2014), Mohlke, K.L. & Boehnke, M. (2015) and Zhao, W. et al. (2017) were analysed [1-5]. A group of potential *loci* of interest with reported association to glycaemic traits and T2D predisposition was enumerated.

The obtained list of genes was further annotated with epigenetic markers of chromatin state (H3K4me1, H3K4me3, H3K27ac) and binding of CTCF proteins and specific pancreatic TFs (PDX1, FOXA2, NKX2.2, MAFB and NKX6.1) using UCSC Genome Browser (GRCh36/hg18 /Human) (http://genome-euro.ucsc.edu/), retrieving data from Pasquali and colleagues, from the ENCODE project [4, 6, 7] (http://genome-euro.ucsc.edu/ENCODE/cellTypes.html) and from the Islet Regulome Browser (GRCh37/hg19/Human) (http://www.isletregulome.org/isletregulome/).

The Zebrafish Information Network (ZFIN) (https://zfin.org/) was employed to search for zebrafish orthologues of the candidate genes [104]. UCSC Genome Browser was used to explore BAC clone libraries retrieved from NCBI Clone DB database [105].

2. Zebrafish maintenance and microinjection

a) Zebrafish facility and husbandry

Zebrafish adults were maintained in the i3S zebrafish facility, under controlled abiotic and biotic parameters, in a recirculating housing system. Adult animals in the facility are kept in 3,5 L water tanks, at a density of 5 adults/L, and they are fed two to three times a day. Zebrafish larvae are kept in smaller tanks at a density of 40 larvae/L. The photoperiod is defined by a daily cycle of 14 hours of light and 10 hours of darkness. Several water parameters are automatically regulated by a WTU equipment – water temperature (26-27ºC), conductivity (0-5g/L) and pH (6,8-8,5). The recirculating housing system includes mechanic, chemical and biological filters.

Rearing, housing and experiments on zebrafish animals were conducted following ethical guidelines and minimizing animal stress and suffering [53]. The i3S animal facility and the research project are licensed by the *Direcção Geral de Alimentação e Veterinária* (DGAV). Protocols followed were approved by the i3S Animal Welfare and Ethics Review Body.

b) Zebrafish breeding and embryos collection

Crossing of adult zebrafish was conducted on breeding cages containing an interior cage with a bottom mesh. During late afternoon, males and females were placed in breeding cages, in a proportion of 1:2, respectively, and separated by a partition. In the morning of the day after, the partition was removed. The breeding tank was placed in a slanted position and under direct light, to instigate reproduction. After spawning, fertilized eggs were collected by filtering the water in each cage with a net. Embryo collection was completed within 20 minutes (min) after partition removal, in order to collect batches of synchronized embryos at the 1-cell-stage of development (used for microinjection).

Zebrafish embryos were maintained in E3 medium – 5 mM NaCl, 0,17 mM KCl, 0,33 mM $CaCl_2 \cdot 2H_2O$, 0,33mM $MgSO_4 \cdot 7H_2O$ and 0,01 % methylene blue (BioChemica, #C.I. 52015). E3 medium was supplemented with 0.01 % (w/v, weight/volume) PTU (1-phenyl-2-thiourea) in cases when pigmentation inhibition of the embryos was required.

c) Embryos bleaching and rear of embryos

Embryos collected from crossing sets in the quarantine room were bleached between 10 and 28 hpf. First, a bleach solution at 0,0036 % was prepared, by the diluting 360 uL of sodium hypochlorite in 1L of distilled water. After that, five washing containers were placed in the following order: bleaching solution, tap water, bleaching solution, and two more of tap water. During this process, embryos collected in a tea strainer were incubated in each bath 5 min. Then, they were transferred to a new Petri dish with E3 medium and incubated at 28ºC.

After a period of 5 to 7 days, zebrafish reached the larval developmental stage and were placed into the zebrafish facility nursery, where they were fed three times a day. Around 3 months after, fish were moved to adult tanks.

d) Microinjection in one-cell stage zebrafish embryos

Microinjection of zebrafish embryos was performed using the Narishige IM-300 Microinjector (Tritech Research). Glass capillaries were pulled using Narishige PN-31 Horizontal Needle Puller (Tritech Research) to generate two glass needles. After the tip was cut, the needle was placed in the microinjector and it was filled with the microinjection solution. The microinjection solution was previously prepared, containing plasmid DNA (concentration dependent on the experiment), *Tol2* transposase mRNA at

a final concentration of 25 ng/uL and 0,05% of phenol red. Zebrafish fertilized embryos, collected as described in Section 2 b), were aligned in Petri plates and injected while in one-cell stage. Each embryo was injected with approximately 2-5 nL of microinjection solution, estimated by the size of the injected drop. Upon microinjection, embryos were incubated at 28,5 ºC in E3 medium with 0,01 % (w/v) PTU.

3. *Tol2* transposase mRNA synthesis

a) *Tol2* transposase mRNA transcription *in vitro*

The *Tol2* transposase was synthetized from the plasmid Tol2-pCS2FA, containing Tol2 cDNA (complementary DNA) [106]. The plasmid was transformed into Mach1™ *E. coli* chemically competent bacteria (described in Section 6 a)) and selected in LB agar plates containing ampicillin (100 µg/mL, Sigma-Aldrich, #A1593). Isolated colonies were picked from the plate to LB liquid medium containing 100 µg/mL of ampicillin and the bacterial cultures grew O.N. at 37 °C, with shaking (220 rpm). The plasmid was transformed into chemically competent bacteria – described in Section 6 a). Plasmid DNA extraction was performed using NZYMiniprep kit (NZYtech, #MB01002), according to manufacturer's instructions. After DNA extraction, the plasmid was linearized by digestion O.N. with *NotI* restriction enzyme (Anza™, #ER0591 – Thermo Fisher Scientific) at 37 °C. Reaction mix included 5 uL of plasmid DNA (approximately at 1200 ng/uL), 1 uL of *NotI* restriction enzyme (10 U/uL), 2 uL of 10x Anza™ Buffer and nuclease-free water up to a final volume of 20 uL. Digestion products were analysed through electrophoresis in 1% (w/v) in 1x TAE buffer (Tris-acetate-EDTA) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201) and run along with 1 kb DNA Ladder (BIORON). Linearized DNA was purified by phenol/chloroform (Section 5), followed by quantification by spectrophotometric analysis.

The linearized and purified Tol2-pCS2FA plasmid was used as template for *Tol2* transposase *in vitro* transcription. All incubations mentioned hereafter were made at 37 °C. Transcription reaction mixes were prepared comprising 6,5 uL of HyPure H_2O (HyClone™ Water, GE Healthcare – Thermo Fisher Scientific), 10 uL of 5x transcription buffer, 5 uL of NTP mix (10 mM A, 10 mM U, 10 mM C, 5 mM G) and 5 uL of DTT (50 mM Dithiothreitol, NZYTech, #MB03101). After an incubation of 5 min, 5 uL of 5' CAP (20-25 mM G(5')ppp(5')G RNA Cap Structure Analog; NewEngBiol, #S1407S) were added and the reaction was incubated for 1 min. Then, 12 uL of purified Tol2-pCS2FA plasmid and 2,5 uL of NZY Ribonuclease Inhibitor (NZYTech, #MB08405) were

sequentially added, each followed by 1-min incubations. At this point, 2 uL of RNA polymerase SP6 (20 U/µL, ThermoFisher Scientific™, #EP0131) were added and the reaction was incubated for 1 hour. Upon incubation, this step was repeated by adding more 1 uL of RNA polymerase. Finally, the template DNA was digested by adding 2 uL of DNase and incubation for 1 hour.

b) Purification of *Tol2* transposase mRNA

mRNA obtained by *in vitro* transcription was purified using ProbeQuant™ G-50 Micro Columns kit (Sigma-Aldrich, GE Healthcare, #GE28-9034-08), following an adapted protocol for purification of radiolabelled probes. Starting on the column preparation, the resin was resuspended by vortexing. Then, the cap was loosened one-quarter turn, the bottom closure was removed and the column was placed in the collection tube, followed by centrifugation at 735 g for 1 min (VWR Micro Star 17, #521-1646). Immediately after, the column was placed in a new RNAse free 1,5 mL Eppendorf and 50 uL of the newly synthesized RNA were slowly loaded in the column, to the top-center of the resin. Upon careful loading, the sample was eluted through centrifugation at 735 g for 2 min. The RNA sample obtained was then purified by phenol/chloroform (Section 5), followed by quantification by spectrophotometric analysis and storing at -80 ºC.

4. DNA extraction from zebrafish embryos and small fish

Zebrafish genomic DNA (gDNA) was extracted from whole embryos, after collection and dechorionation, or small fish. First, embryos were placed in 40 uL of CHELEX solution. The CHELEX solution consists in 5% Chelex® 100 sodium form resin (Sigma-Aldrich, #C7901-25G) diluted in TE buffer (10 mM Tris·Cl pH 8,0; 1 mM EDTA). Then, 5 uL of proteinase K (10 mg/mL) were added for each embryo into solution and this was incubated O.N. at 56 ºC with shaking. After that, the proteinase K was inactivated by incubation at 100 ºC for 10 min, followed by vortexing. Before pipetting gDNA to be used in PCR reactions, the samples were briefly centrifuged. Extractions of gDNA were stored at -20ºC.

5. Phenol/Chloroform purification

Preparations of DNA and RNA were purified with Phenol/Chloroform. Eluted DNA/RNA preparations extracted with NZYMiniprep kit (NZYtech) were adjusted to a

final volume of 100 uL with HyPure H_2O (HyClone™ Water, GE Healthcare – Thermo Fisher Scientific), followed by adding 100 uL of Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v, Invitrogen™, #P3803 – Thermo Fisher Scientific). The mixture was vortexed and centrifuged at 13000 rpm for 5 min (VWR Micro Star 17, #521-1646). Upon centrifugation, the aqueous phase was transferred to a new RNAse free 1,5 mL Eppendorf and mixed with 100 uL of chloroform (Fisher Chemicals). After vortexing, this mixture was centrifuged in the same conditions mentioned above. The aqueous phase was again collected to a new RNAse free 1,5 mL Eppendorf. DNA/RNA precipitation was achieved by adding of 10 uL of Sodium Acetate (AcNa, 3 M, pH 5,6) and 200 uL of ice-cold ethanol (EtOH, 100 %) per 100 uL of aqueous phase collected, followed by incubation of the sample at -20 ºC for 1 hour (minimum time period). The sample was then centrifuged at 13000 rpm for 15 min at 4 ºC (VWR Micro Star 17R, #521-1647) and the supernatant was discarded. DNA/RNA was washed with ice-cold EtOH 70 %, followed by a centrifugation at 13000 rpm for 5 min and removal of the supernatant. Pellet was dried out resuspended in 15 uL DEPC-treated H_2O.

Purified DNA/RNA was quantified by spectrophotometric analysis, using NanoDrop 1000 Spectrophotometer (ThermoFisher Scientific™). DNA preparations were stored at -20 ºC and RNA preparations at -80 ºC.

6. Characterization of putative regulatory elements

a) PCR amplification and subcloning of putative *PDX1* CREs

Human sequences of putative enhancers and insulators were tested for activity through *in vivo* assays in zebrafish. Primers flanking the regulatory elements putative sequences were designed (**Table 3**). Sequences were amplified using the proofreading enzyme i-MAX™ II Taq DNA Polymerase (iNtRON Biotechnology, Inc., #25261), suitable for TA cloning. PCR reactions were set up comprising 2 uL of 10x PCR buffer, 2 uL of dNTP mixture (2,5 mM each), 0,4 uL of forward and reverse primers (10 uM each), 1 uL of template DNA (0,1-10 ng/uL), 0,3 uL i-MAX™ II Taq DNA Polymerase (5U/uL) and nuclease-free water up to a final volume of 20 uL. As template DNA, the PDX1 BAC and a commercially available sample of human genomic DNA (gDNA) obtained from female and male anonymous donors were used. PCR amplifications were performed in Veriti thermocycler (Applied Biosystems) – this equipment was used in all PCR protocols – applying the following conditions: initial denaturation step at 94 °C for 3 min, followed by

35 cycles of 94 °C for 30 s, 56-60 °C (depending on the primers melting temperature) for 45 s and 72 °C for 1 min/kb, and one last extension step of 72 °C for 10 min.

Table 3. Primers for PCR amplification of PDX1 putative CREs. Forward (FW, **5'-3'**) and reverse (RV, **3'-5'**) primers, employed melting temperatures and amplicon size of indicated regions.

Primer name	Primers'	Tm (°C)	Amplicon size (bp)
en1 PDX1 FW	GCAGTAAACAGACTCCAGCC	60-62	1676
en1 PDX1 RV	AGATAGTGTGGGGTGGGGG		
en2 PDX1 FW	AGCCTACACCCCTGGACCC	58	1532
en2 PDX1 RV	GCAGCCTCCATGTTCTCTTGG		
eSNP PDX1 FW	GAAAATATTTAAACAACGCCTGGC	58	1159
eSNP PDX1 RV	TGGTATCCAGGTCTGAGAGG		
ins1 PDX1 FW	CTTAGAAATGCCCCTGCTATGC	60	1476
ins1 PDX1 RV	GACGCATCTGATGCCAACTGG		
ins2 PDX1 FW	CTTGTTGGGAAAAAAAGTCTCCC	56	795
ins2 PDX1 RV	CAAATCATCCTGGGAAAAAGTAGC		
ins3 PDX1 FW	AGTTTTTCCTGTAGGTGGGCT	56	1143
ins3 PDX1 RV	AATGCCATCAGACACCTGTGA		

The PCR amplification was confirmed by running of the PCR product in an 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON). Loaded agarose gels were visualized by transillumination under UV (TFX – 35 M, VILBER LOURMAT). In cases where the PCR product was not specific, the fragment of the correct size was extracted from the agarose gel and purified with the NZYGelpure (NZYtech, #MB01102), according to the standard protocol.

Purified PCR products were TA cloned into the entry vector pCR™8/GW/TOPO® (Invitrogen™, #250020 – Thermo Fisher Scientific). TOPO® cloning reactions were set up according to the standard protocol, comprising 4 uL of fresh PCR product. The reaction was incubated for 30 min at room temperature (RT). The cloning reaction was followed by transformation of Mach1™ *E. coli* chemically competent bacteria. A total of 3 uL of cloning reaction was incubated with 50 uL of competent bacteria for 30 min on ice. Cells were exposed to a heat-shock for 30 s at 42 °C, and immediately transferred to ice for 2 min. Cells recovery was achieved by addition of 400 uL of Lysogeny Broth (LB) and one-hour incubation at 37 °C, with shaking (220 rpm). Cells were plated on LB agar plates containing 100 µg/mL of spectinomycin antibiotic (Sigma-Aldrich, #S0692) and incubated overnight (O.N.) at 37 °C. Isolated colonies were picked from the plate to a liquid medium of 3 mL of LB containing 100 µg/mL of spectinomycin. Liquid cultures

grew O.N. at 37 °C, with shaking (220 rpm). Plasmid DNA extraction was performed using NZYMiniprep kit (NZYtech), according to manufacturer's instructions, and stored at -20 ºC.

Presence of the correct fragment into pCR™8/GW/TOPO® plasmid was tested through a diagnostic enzymatic reaction with *EcoRI* (Anza™, #IVGN0116 – Thermo Fisher Scientific). Enzymatic reactions included 2 uL of plasmid DNA (200-400 ng/uL), 0,3 uL of *EcoRI* restriction enzyme (20 U/uL), 2 uL of 10x Anza™ Buffer and nuclease-free water up to a final volume of 20 uL. Digestion reactions occurred for 2 hours at 37 °C and the digestion products were analysed through electrophoresis in an 1% (w/v) in 1x TAE agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), run along with 1 kb DNA Ladder (BIORON). Loaded agarose gels were visualized by transillumination under UV (TFX – 35 M, VILBER LOURMAT). Validation of the correct fragment was further confirmed by Sanger sequencing using the following primers: M13 forward (5'-TGTAAAACGACGGCCAGT-3') and M13 reverse (5'-TCAGGAAACAGCTATGAC-3'). Sequencing reactions were carried out by the in-house Genomics Core Facility – GenCore. Sequence alignment of the results was performed on Benchling ([Biology Software]. (2019). https://benchling.com).

b) Recombineering from pCR™8/GW/TOPO® into destination vector

The putative enhancer and insulator sequences subcloned into pCR™8/GW/TOPO® plasmid were recombined into a destination vector, using a Gateway® recombination strategy. Putative enhancers were recombined into the Z48 vector [39] and putative insulators were recombined into the insulator test vector [15].

Recombination was performed with Gateway® LR Clonase® II Enzyme mix (Invitrogen™, #11791020 – Thermo Fisher Scientific). The recombination reaction was set up to a final volume of 2,5 µL, containing 1 µL of entry vector (50 ng/µL), 1 µL of destination vector (50 ng/µL) and 0,5 µL of Clonase® II Enzyme mix. Reaction mixes were incubated in the thermocycler (Veriti, Applied Biosystems) for 1 hour at 25 ºC. The reaction was stopped upon adding of 0,25 µL of Proteinase K solution (2 µg/µl) and incubation for 10 min at 37ºC. The product of recombination was used to transform Mach1™ *E. coli* chemically competent bacteria, as it is described in section 5 a). Cells were plated on LB agar plates containing 100 µg/mL of ampicillin antibiotic (Sigma-Aldrich, #A1593) and incubated O.N. at 37 °C. Isolated colonies were picked from the plate to a liquid medium of 3 mL of LB containing 100 µg/mL of ampicillin. Liquid cultures

grew O.N. at 37 °C, with shaking (220 rpm). Plasmid DNA extraction was performed using NZYMiniprep kit (NZYtech), according to manufacturer's instructions, and stored at -20 ºC.

(1) Z48 vector recombination

The Z48 vector (**Figure 3A**) contains a Green Fluorescent Protein (GFP) reporter gene under the control of a minimal promoter. The Z48 enhancer, cloned upstream the minimal promoter, drives GFP expression in zebrafish midbrain, working as internal control of transgenesis [107]. A Gateway® site located between the Z48 enhancer and the minimal promoter allows to recombine the putative enhancer sequence from pCR™8/GW/TOPO® to this destination vector.

(2) Insulator test vector recombination

The insulator test vector (**Figure 3B**) contains a GFP reporter gene under a Cardiac Actin promoter, which drives GFP expression in muscle cells. The Z48 enhancer is located upstream of a Gateway® site, that is followed by the Cardiac Actin promoter. The Gateway® site allows to recombine the putative insulator sequence from pCR™8/GW/TOPO® (entry vector) to this vector. After microinjection in one-cell stage zebrafish embryos, insulator activity can be accessed through compared analysis of Z48-mediated GFP expression in the midbrain and GFP expression in muscle cells.

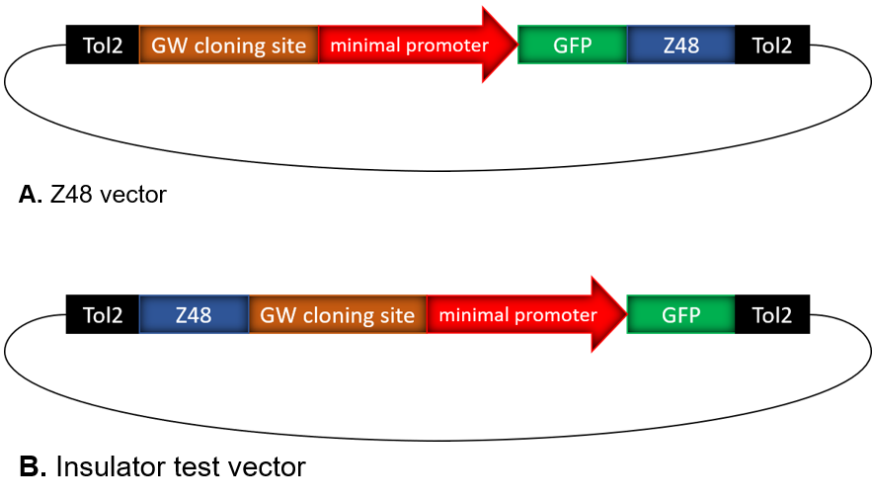

Figure 3. Vectors for detection of cis-regulatory elements. The vectors are transposable elements in the presence of *Tol2* transposase [102], as they contain *Tol2* recognition sites flanking the region of interest. Each vector contains a minimal promoter, the Cardiac Actin promoter, driving expression of GFP in the somites of transgenic embryos. The transposons also include a midbrain-specific enhancer, the Z48 enhancer, which drives strong GFP expression in the midbrain. A. The Z48 vector contains a Gateway® cloning site upstream the minimal promoter. When the vector is empty, GFP expression is detected in midbrain and somites. The insertion of an active tissue-specific enhancer in the cloning site leads to GFP expression in that tissue. B. The insulator test vector comprises the Z48 enhancer upstream the GW cloning

site, followed by the minimal promoter. The empty vector induces a similar GFP expression pattern to the empty Z48 vector. Upon cloning of a strong insulator, GFP expression in the midbrain is significantly decreased, due to blocking of activity by the Z48 enhancer over the promoter.

c) Microinjection of vectors for detection of cis-regulatory elements in zebrafish embryos

The Z48 vector containing the putative enhancers and the insulator test vector containing the putative insulators were microinjected in one-cell stage zebrafish embryos, as described in Section 2 d). In this assay, the following zebrafish lines used were the following: the transgenic line tg(sst:mCherry) for testing of putative enhancers and the wt animals for testing of putative insulators. Upon microinjection along with *Tol2* transposase mRNA, the transposable element flanked by the *Tol2* recognition sites is integrated into the zebrafish genome.

Each microinjection solution was prepared containing the plasmid DNA and the *Tol2* transposase mRNA, both at a final concentration of 25 ng/uL, and 0,05% of phenol red. After microinjection, embryos were incubated at 28,5 ºC in E3 medium supplemented with 0,01 % (w/v) PTU.

7. Immunohistochemistry and immunostaining of zebrafish embryos

Enhancer activity was evaluated by immunohistochemistry for pancreatic markers. Embryos at a developmental stage of 36 hpf were dechorionated and fixed O.N. at 4 ºC in formaldehyde 4 % in PBS – Phosphate-Buffered Saline, 1x. Upon fixation, all washes and incubations were done with shaking. Fixed embryos were washed in 0,1 % Triton X-100 in PBS-1x (PBS-T) for 5 min at RT, followed by a 2 hours permeabilization with 1 % PBS-T at RT. Then, embryos were washed once more in 0,1 % PBS-T (0,1 % Triton X-100 in PBS-1x) for 5 min at RT and blocking was performed with 5% BSA-PBS-T (5% Bovine Serum Albumin in 0,1% PBS-T) for 1 hour at RT. After the blocking, immunohistochemistry was performed by incubation of the embryos with anti-Nkx6.1 (1:75, Hybridoma Bank, #F55A10) primary antibody diluted in 5% BSA-PBS-T, throughout 36 hours at 4 ºC. This step was followed by 6 washes of 10 min each in 0,1 % PBS-T at RT. Then, the embryos were incubated with DAPI (1:1000, Invitrogen, #D1306) and the anti-Mouse Alexa Fluor® 647 (1:800, Invitrogen, #A21236) secondary antibody diluted in 5% BSA-PBS-T (O.N. at 4 ºC). Finally, embryos were washed at RT for periods of 10 min in 0,1 % PBS-T at RT, followed by a final wash of 30 min. Upon

removal of the washing solution, the embryos were stored in 50 % glycerol in PBS-1x at 4 ºC.

Microscopy slides were prepared by removing the yolk of the embryos and mounting them on 50 % glycerol in PBS-1x. Imaging analysis was performed in a SP5II Leica confocal microscope. Confocal images were processed with ImageJ software [108].

8. Fluorescence quantification of zebrafish embryos *in vivo*

Embryos with 24 hpf were dechorionated. During the documentation process, embryos were anaesthetized by adding 100-200 mg/L of MS-222 anaesthetic agent, also termed tricaine, (ethyl 3-aminobenzoate methane-sulfonate salt, Sigma-Aldrich, #A5040) to the E3 medium where the embryos were incubated. Between 10 and 30 zebrafish embryos were analysed and photographed in a Leica M205 stereomicroscope. Photographs were taken in 3 % agarose plates, by dilution of agarose in E3 medium. Image analysis and fluorescence quantification was performed using the Fiji software from ImageJ [109].

9. Humanization of the zebrafish genome

a) BAC clone extraction and confirmation

BAC clone libraries retrieved from NCBI Clone DB database were explored in UCSC Genome Browser (GRCh38/hg38/Human) (http://genome-euro.ucsc.edu/), as it was mentioned in Section 1 [7]. BAC clone CH17-423D7 from CHORI-17: Hydatidiform Mole (*Homo sapiens*) BAC Library was selected (position chr13:27,763,911-27,985,654, GRCh38/hg38/Human). The clone was purchased online from BACPAC Resources Center (BPRC) (https://bacpacresources.org/), at Children's Hospital Oakland Research Institute in Oakland, California, USA. BAC clones from the hydatidiform mole were created at BACPAC Resources by Drs. Mikhail Nefedov & Pieter J. de Jong using a cell line created by Dr. Urvashi Surti.

The CH17-423D7 BAC (henceforward referred as PDX1 BAC) was extracted from the original bacterial strain sent from BPRC. High-quality BAC DNA was prepared and purified using NucleoBond® BAC 100 kit (#740579, Macherey-Nagel GmbH & Co. KG), following manufacturer's instructions. Purified BAC DNA was quantified by spectrophotometric analysis, using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific).

BAD DNA integrity was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), run along with Lambda DNA/HindIII Marker (Thermo Scientific™, #SM0101).

Validation of the correct human genomic sequence in the BAC was confirmed by PCR amplification with primers flanking the putative regulatory elements sequences selected before (**Table 6**). Sequences were amplified using NZYTaq II 2× Green Master Mix (NZYtech, #MB35802). PCR reactions were set up accordingly to the standard PCR mix, including 0,1 uL BAC template (10 ng/uL). PCR conditions were as follows: initial denaturation step at 95 °C for 3 min; 35 cycles of 94 °C for 30 s, 56-60 °C (depending on the primers melting temperature) for 30 s and 72 °C for 1 minute/kb; and final extension step of 72 °C for 5 min. PCR amplification was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON).

b) PDX1 BAC recombineering

(1) BAC electroporation

In order to perform recombineering, PDX1 BAC was transformed into the recombinogenic bacteria SW102, purchased from National Cancer Institute (NCI-Frederick) [110]. The bacterial glycerol stock of *E. coli* SW102 was streaked in a LB agar plate and incubated O.N. at 32°C. Cells were then streaked to LB agar plates containing 12,5 µg/mL of tetracycline antibiotic (Sigma-Aldrich, #87128) and incubated O.N. at 32°C. Isolated colonies were picked from the plate to LB liquid medium containing 12,5 µg/mL of tetracycline. Liquid cultures grew O.N. at 32 °C, with shaking (220 rpm) and isolated from direct light – due to antibiotic light sensitivity. Liquid cultures were used to prepare glycerol stocks with sterile 25 % glycerol diluted in double-distilled water (ddH_2O) (Sigma-Aldrich, #G9012).

Electrocompetent *E. coli* SW102 cells were prepared from the above-mentioned liquid cultures. After O.N. growth at 32 °C, 1 mL of liquid culture was diluted in 100 mL LB containing 12,5 µg/mL of tetracycline (1:100). Cells were grown at 32 °C to an OD_{600nm} of 0,4-0,5, followed by cooling on ice-water bath for 20 min. Cells were kept on ice until electroporation during the following steps. Cells were transferred to two pre-chilled 50 mL Falcon tubes (BD Biosciences) and harvested by centrifugation at 4000 rpm for 10 min at 4 °C (Refrigerated Centrifuge Sigma 3k15 – Sigma-Aldrich). Upon centrifugation, the supernatant was removed and each pellet was resuspended in 50 mL of 10 % ice-

cold glycerol. Three washing cycles with centrifugation and resuspension on glycerol were followed, with increasing centrifugation speeds (5000 rpm, 6500 rpm and 8000rpm sequentially). The supernatant was decanted leaving around 1 mL of glycerol, where cells were gently resuspended by stirring and kept at 4 °C until electroporation.

An aliquot of 40 uL of electrocompetent cells was used for each electroporation in a pre-chilled 0,1 cm cuvette (Bio-Rad, #1652089). Cells were previously mixed with purified plasmid DNA and incubated for 5-10 min at 4 °C. Plasmid DNA concentrations tested ranged between 200 ng and 1 ug, in a final volume of 1 to 5 uL. Electroporation was performed in the Gene Pulser® Electroporation System (Bio-Rad) and the conditions tested were the following: 25 mF, 1,2-1,8 kV and 200 Ω. After electroporation, cells were recovered by adding of 1 mL of pre-warmed LB, transferring to a 15 ml Falcon tube and incubation at 32 °C for 2 hours with shaking (220 rpm). Cells were plated on LB-chloramphenicol agar plates (12,5 ug/mL) and incubated O.N. at 32 °C. A known amount of supercoiled pD274 (2,8 kb) plasmid was transformed, followed by cells recovery and plating on LB agar plates containing 100 ug/mL of kanamycin antibiotic (Sigma-Aldrich, #BP861). The resulting colonies were counted in order to assess the efficiency of electrocompetent bacteria, calculated as number of transformed bacteria per microgram of DNA.

Bacterial clones grown on selective plates with chloramphenicol were confirmed for the presence of the BAC by colony PCR. The set of primers designed to amplify the putative regulatory elements selected above were used (**Table 6**). PCR amplification was performed using NZYTaq II 2× Green Master Mix (NZYtech, #MB35802). Reactions were prepared accordingly to the standard PCR mix, followed by the picking of single colonies to each PCR reaction. Amplification was performed with the following conditions: initial denaturation step at 95 °C for 3 min; 35 cycles of 94 °C for 30 s, 56-60 °C for 30 s and 72 °C for 1 minute/kb; and final extension step of 72 °C for 5 min. PCR amplification was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON).

Additionally, single colonies of *E. coli* SW102 cells containing the BAC (SW102:BAC) were tested through antibiotic selection in LB agar plates containing both chloramphenicol and tetracycline antibiotics (12,5 µg/mL and 100 µg/mL, respectively).

(2) PCR amplification of Tol2 cassette

Taking advantage of the recombinase functions of the *E. coli* SW102 strain, a cassette containing the minimal sequences required for *Tol2* transposition in zebrafish

was placed inside the PDX1 BAC. The cassette includes an ampicillin resistance gene flanked by the *Tol2* recognition sequences, allowing antibiotic selection. Primers to amplify the Tol2 cassette were designed, each comprising a 5' sequence of 50 base pairs (bps) with homology to the target site on the backbone of the BAC – an *attB1* site. Primers used were the following: 5'-cgtaagcggggcacatttcattacctctttctccgca cccgacatagataCCCTGCTCGAGCCGGGCCCAAGTG-3' and 5'-cggggcatgactattggcg cgccggatcgatccttaattaagtctactagATTATGATCCTCTAGATCAGATCT-3'. The plasmid pCR8GW-iTol2, kindly sent by Professor K. Kawakami, was used as template for PCR amplification. The Tol2 cassette was amplified using the proofreading enzyme i-MAX™ II Taq DNA Polymerase (iNtRON Biotechnology, Inc., #25261). PCR reactions were set up comprising 10 uL of 10x PCR buffer, 8 uL of dNTP mixture (2.5mM each), 5 uL of forward and reverse primers (10 uM each), 1 uL of template DNA (10 pg/uL), 0,7 uL i-MAX™ II Taq DNA Polymerase (5U/uL) and nuclease-free water up to a final volume of 100 uL. PCR amplifications were performed applying the following conditions: an initial denaturation step at 94 °C for 2 min, followed by 35 cycles of 94 °C for 30 s , 58 °C for 30 s and 72 °C for 1 minute/kb, and one last extension step of 72 °C for 5 min. PCR amplification was confirmed by running 1/10 of the PCR product in an 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON). Complete digestion of the template plasmid was ensured by adding 2 uL of *DpnI* restriction enzyme to the PCR reaction and an O.N. incubation at 37 °C. At last, the PCR product was purified using the NZYGelpure (NZYtech, #MB01102), according to the standard protocol. The purified Tol2 cassette was quantified by spectrophotometric analysis, using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific).

(3) BAC recombineering of the Tol2 cassette

The term "recombineering" stands for recombination-mediated genetic engineering, referring to an *in vivo* technique of genetic manipulation [63, 64]. This method allows modifying DNA sequences via homologous recombination performed by bacteria, which displays several advantages in comparison to classical cloning methods, when necessary to manipulate large sequences [61]. *E. coli* SW102 genome contains a Red recombinase system with three λ Red-encoded genes, regulated by a heat-shock inducible promoter. The three genes comprise *gam*, *bet* and *exo* – *gam* encodes a protein responsible for preventing the destruction of the exogenous DNA fragments by endonucleases, *bet* encodes a protein that promotes the annealing between complementary single-stranded DNA sequences and *exo* encodes an exonuclease with

5' to 3' activity able to produce 3' overhangs, enabling recombination of homologous fragments [63, 64]. This system induces recombinase functions after transformation of bacteria containing a BAC DNA (in this case, PDX1 BAC) with a donor DNA sequence flanked by homologous sequences to the target site of recombineering.

To perform recombineering, electrocompetent *E. coli* SW102:BAC cells were prepared from the positive single colonies selected in section 9 b) (2). Isolated colonies were picked from the plate to LB medium containing chloramphenicol and tetracycline (12,5 µg/mL and 100 µg/mL, respectively) and liquid cultures grew O.N. at 32 °C, with shaking (220 rpm) and isolated from direct light. The day after, 1 mL of O.N. culture was diluted in 100 mL of LB-chloramphenicol-tetracycline (12,5 µg/mL and 100 µg/mL, respectively). Cells were grown at 32 °C to an OD_{600nm} of 0,4-0,5, followed by a heat-shock at 42 °C for exactly 15 min, with manual shaking, to induce recombinase functions. Upon heat-shock, cells were immediately transferred to an ice/water slush and left to cool for 10 min. Cells were treated accordingly to the protocol mentioned in section 9 b) (1). After the last centrifugation step, cells were resuspended in 1 mL of 10 % glycerol and kept at 4 °C until electroporation.

Aliquots of 40 uL of freshly-prepared electrocompetent cells were mixed with the purified Tol2 cassette (3,5–700 ng), along with the competence control test. Electroporation was performed as mentioned in section 9 b) (1). Cells were plated on LB agar selective plates, containing chloramphenicol and ampicillin (12,5 µg/mL and 100 µg/mL, respectively), and incubated O.N. at 32 °C.

(4) Selection and preparation of BAC-Tol2 construct

Single colonies of *E. coli* SW102 cells grown on LB-chloramphenicol-ampicillin plates (12,5 µg/mL and 100 µg/mL, respectively) were confirmed for recombineering through colony PCR and antibiotic resistance.

Colony PCR was performed with two pairs of primers. Two primers were designed flanking the target site on the BAC where the Tol2 cassette is placed in case of recombination (BAC REC primers for colony PCR A). Additionally, a forward primer was designed to the ampicillin resistance gene of the Tol2 cassette and used together with the previously mentioned reverse primer on the homology region. PCR amplification was performed using NZYTaq II 2× Green Master Mix (NZYtech, #MB35802). Reactions were set up comprising the standard PCR mix, followed by picking of single colonies. Amplification was performed with the following PCR conditions: initial denaturation step at 95 °C for 3 min; 35 cycles of 94 °C for 30 s, 60 or 65 °C for 30 s and 72 °C for 1 min/kb;

and final extension at 72 °C for 5 min. PCR amplification was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON).

Antibiotic selection of *E. coli* SW102 cells containing the BAC-Tol2 construct (SW102:BAC-Tol2) was achieved by streaking of single colonies on LB-chloramphenicol-tetracycline (12,5 µg/mL each) and LB- chloramphenicol-tetracycline-ampicillin (12,5 µg/mL, 12,5 µg/mL and 100 µg/mL, respectively) agar plates. Incidence of false positives was also verified by streak of single colonies on LB-spectinomycin (100 µg/mL) agar plates, the resistance present in the backbone of the plasmid used as template for the amplification of the Tol2 cassette.

Upon selection of positive colonies for Tol2 cassette recombination into the BAC, plasmid DNA was extracted and purified using NucleoBond® BAC 100 kit (#740579, Macherey-Nagel GmbH & Co. KG), following manufacturer's instructions. Quantification was performed by spectrophotometric analysis, using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). The integrity of the BAC-Tol2 construct was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), run along with Lambda DNA/HindIII Marker (Thermo Scientific™, #SM0101). Furthermore, the construct was purified by Phenol/Chloroform, as it is described in Section 5.

c) BAC transgenesis in zebrafish

(1) Zebrafish *pdx1* mutant: genotyping and expanding the line

A zebrafish mutant line for the *pdx1 locus* was reared in the husbandry of the i3S zebrafish facility, according to ethical guidelines and standard protocols [53]. A zebrafish mutant line holding *pdx1* mutant allele designated pdx1$^{sa280/sa280}$, generated through the Zebrafish Mutation Project, was kindly sent by Professor Robin A. Kimmel [111]. The zebrafish line was characterised by Kimmel, R. A. and colleagues [93]. The *pdx1* mutant allele yields a transcript containing a premature stop at codon (Y37X), located within the protein highly conserved N-terminal transactivation domain.

The offspring of the incross between two heterozygous animals for the pdx1^{sa280} allele was reared in the husbandry of the facility. When the offspring reached sexual maturity, animals were genotyped through an outcross with wild-type (wt) animals. The offspring of this outcross was then genotyped by DNA extraction of 8 embryos batches (Section 4), following Kimmel, R. A. and colleagues' recommendations [93]. The *pdx1*

locus was amplified with the following primers: forward 5'-CCCCAACGAAGACTACAGCC-3' and 5'-ATGGCCTGCAATCAGGAGTTA-3'. The PCR reaction was performed using NZYTaq II 2× Green Master Mix (NZYtech, #MB35802), preparing the standard PCR mix. PCR conditions were the following: 95 °C for 3 min; 35 cycles of 94 °C for 30 s, 61 °C for 30 s and 72 °C for 30 s; and 72 °C for 5 min. After confirming the amplification on a 1% (w/v) agarose gel, the PCR product was digested with *DraI* restriction enzyme (Thermo Fisher Scientific™, #ER0221). Enzymatic reactions included 3 uL of PCR product, 1 uL of *DraI* (10 U/uL), 2 uL of 10x Buffer Tango and nuclease-free water up to a final volume of 20 uL. A negative control for the enzymatic reaction was performed along, preparing the same reaction but without *DraI* restriction enzyme. Digestion reactions occurred O.N. at 37 °C and the final product was analysed in an 2 % (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), run along with 1 kb DNA Ladder (BIORON).

(2) BAC microinjection and viability assays

The purified BAC-Tol2 construct was microinjected in one-cell stage zebrafish embryos, as described in Section 2 d). In this assay, the following zebrafish lines used were following: the *pdx1* mutant line with the sa280 allele and the WT line from the i3S facility.

Each microinjection solution was prepared containing the plasmid DNA at final concentration ranging from 50 to 500 ng/uL, *Tol2* transposase mRNA at a final concentration of 25 ng/uL and 0,05% of phenol red. After microinjection, embryos were incubated at 28,5 ºC in E3 medium.

Zebrafish embryos microinjected with the BAC-Tol2 construct were analysed at 24 and 48 hpf to assess viability. The surviving embryos were bleached as described in Section 2 c) and reared in the i3S zebrafish facility (ongoing).

(3) Genotyping BAC transgenesis

Samples of gDNA extracted from batches of four zebrafish individuals with 10 days post-fertilization (dpf) microinjected with the *PDX1* BAC were used for PCR amplification. Primers designed to bind specifically to the promoter of the human *PDX1 locus* were used to evaluate integration of the human transposon into the zebrafish genome. Primers used were the following: BAC genotyping forward (5'-AGCCTCTGCTTCAGCTTCTG-3') and BAC genotyping reverse (5'-GGTGCAGAAACAAGCCTCTC-3').

As controls, PCR reactions containing zebrafish DNA mixed with the PDX1 BAC were performed. To define a minimum threshold in which the human *PDX1* locus could

be detected within zebrafish genome, the amount of BAC DNA employed in PCR reactions was estimated to mimic the event of a single molecule insertion of the human PDX1 BAC into the zebrafish genome. Taking into account that the full sizes of the zebrafish genome and the PDX1 BAC molecule are ~1,7x10^9 and ~3x10^5 bp, respectively [105, 112], the zebrafish genome is approximately 10^4 larger than the BAC molecule. Therefore, PCR reactions containing mixtures of zebrafish and BAC DNA were prepared in a proportion of 1:10000. Considering an approximated DNA concentration of 10 ng/uL in zebrafish DNA extraction, each reaction was prepared mixing 1 uL of zebrafish DNA with 1 pg of BAC. Further dilutions of the same zebrafish DNA sample with 10 fg and 1 fg of BAC were also prepared to ensure sufficient resolution to mimic the presence of 1 BAC copy in 100 and 1000 zebrafish genome copies, respectively. Mixes are later addressed as “1 BAC + 1 ZF”, “1 BAC + 100 ZF” and “1 BAC + 1000 ZF”.

PCR amplification was performed using NZYTaq II 2× Green Master Mix (NZYtech, #MB35802). Reactions were prepared accordingly to the standard PCR mix and amplification was performed with the following conditions: initial denaturation step at 95 °C for 3 min; 35 or 40 cycles of 94 °C for 30 s, 56-60 °C for 30 s and 72 °C for 1 minute/kb; and final extension step of 72 °C for 5 min. PCR amplification was confirmed by electrophoresis in 1% (w/v) agarose gel stained with GreenSafe Premium (NZYtech, #MB13201), along with 1 kb DNA Ladder (BIORON).

Results and Discussion

1. Candidates selection

To select our *locus* of interest, GWAS datasets that describe human SNPs linked to glycaemic traits and T2D predisposition (see Materials and Methods) were collected, excluding SNPs identified in coding regions. A total of different 126 SNPs potentially impacting on disease susceptibility was listed (**Supplementary Table 1**). Then, a series of guidelines was followed to select a single candidate *locus* from the compiled data.

The first selection criterion was the sorting of SNPs co-localizing with epigenetic marks characterising enhancer elements, thus potentially determining disease susceptibility through cis-regulatory mechanisms. Therefore, we carefully analysed histone modifications – namely H3K4me1 and H3K27ac – detected by ChIP-Seq. H3K4me3 epigenetic mark was also evaluated in order to distinguish regions simultaneously enriched in H3K4 mono- and tri-methylation marks present at gene promoters. Islet-specific TFBS identified both in human pancreatic progenitor cells and adult pancreatic islets were explored in the Islet Regulome browser and their ChIP-Seq tracks analysed [7, 23, 39]. Regarding the TFs screening, it is noteworthy to mention the work of Pasquali et al. in which authors demonstrate that crucial genes regulating pancreatic function of adult lineages as well as pancreas homeostasis are bound by a fraction or all of a cluster of islet TFs. This cluster includes PDX1, MAFB, NKX6.1, FOXA2 and NKX2.2 [7]. This way, screening for their concomitant signature was employed as first criterion along with epigenetic marks.

The second sorting step consisted in assigning every SNP to a specific gene according to genomic proximity, assuming that SNPs potentially modulate gene´s transcription. Moreover, by bibliographic review, all the SNPs already annotated to be associated to differences in expression of a single gene or *locus* (linkage disequilibrium) were included. Analysis of genomic localization of SNPs and nearby genes was performed using UCSC Genome Browser (GRCh37/hg19/Human; see Material and Methods). Genes and associated SNPs were sorted setting an upper limit of 150 kb of genomic distance among the two elements. This analysis aims to ensure that the genomic region of interest could be included in a BAC, that usually have a capacity limit of 300 kb [56]. As we intend to gain knowledge on the impact of human regulatory SNPs on a diabetes-like phenotype through functional studies, we sorted the enumerated SNPs according to their linkage to key-pancreatic genes. Namely, special attention was given to genes previously described to have functional relevance for pancreatic

development and function, such as MODY-associated genes. The previously mentioned series of criterions resulted in a list of 19 candidate *loci* (**Table 4**).

Table 4. List of 19 candidate *loci* defined as potential *loci* of interest. The list enumerates: T2D-associated SNPs retrieved from GWAS datasets; genes located nearby those SNPs, accompanied by number of SNPs per nearby gene; genomic localization coordinates (GRCh37/hg19), length of their coding sequences and bibliographic references for T2D-associated SNPs (A - Sun W. et al. 2018;.B - Mahajan et al. 2018; C - O'Hare et al. 2014; D - Pasquali et al. 2014; E - Mohlke, K.L. et al. 2015; F - Zhao W. et al 2017).

T2D-associated SNP	Nearby gene	Number of SNPs per nearby gene	Genomic localization (GRCh37/hg19)	Length of gene coding sequence (bps)	References
rs7124355	*ABCC8*	3	chr11:17,414,432-17,498,449	84018	A
rs757110					A
rs5219					B
rs7202877	*BCAR1*	2	chr16:75262928-75299905	39867	C
rs72804106					D
rs7163757	*C2CD4A*	2	chr15:62359176-62363116	3941	D
rs7172432					E
rs7945565	*CRY2*	3	chr11:45868957-45904799	35843	D
rs7945689					D
rs1401419					D
rs73300993	*FOXA2*	4	chr20:22561642-22565101	3460	D
rs6048202					D
rs1203898					D
rs1203899					D
rs2908286	*GCK*	2	chr7:44183870-44198887	53900	D
rs4607517					D
rs8108269	*GIPR*	9	chr19:46171502-46185717	14216	B
rs1800437					B
rs11670462					D
rs55872740					D
rs10403962					D
rs10404142					D
rs10404527					D
rs10409882					D
rs8104845					D
rs11651052	*HNF1B*	9	chr17:36046434-36105096	58663	A
rs11263763					A
rs11651755					A
rs11658063					A
rs2005705					A
rs4239217					A
rs4430796					A
rs7501939					A
rs8064454					A
rs4812829	*HNF4A*	2	chr20:42984441-43036115	31590	C
rs1884613					A
rs1002226	*KCNJ11*	3	chr11:17406796-17410206	3411	A
rs2074314					A
rs7124355					A

rs10842994	*KLHL42*	8	chr12:27933187-27955973	22787	E
rs12581729					D
rs10842991					D
rs10771372					D
rs3751239					D
rs10842992					D
rs10842993					D
rs11049161					D
rs11936387	*MAEA*	2	chr4:1283672-1333925	50254	D
rs6815464					E
rs12186664	*PCSK1*	4	chr5:95726040-95768985	42946	D
rs17085593					D
rs59139497					D
rs2882298					D
rs35369009	*PDX1*	1	chr13:28494168-28500451	6284	D
rs10510110	*PLEKHA1*	2	chr10:124134094-124191871	57778	B
rs2421016					E
rs4925115	*SREBF1*	1	chr17:17714663-17740325	25663	C
rs1801214	*WFS1*	8	chr4:6271577-6304992	33416	B
rs1801212					B
rs4689388					D
rs4320200					D
rs13107806					D
rs13127445					D
rs4273545					D
rs6830765					D
rs4457053	*ZBED3*	4	chr5:76383884-76444176	10499	E
rs4457054					D
rs7708285					D
rs7732130					D
rs11634397	*ZFAND6*	3	chr15:80351910-80430735	78826	E
rs1357335					D
rs1357336					D

Further selection ensured that candidates have a zebrafish ortholog and availability of a BAC clone containing the *locus* of interest [104, 105]. Among the 19 candidates (Table 3), we chose *PDX1* as our *locus* of interest. As addressed in the Introduction, *PDX1* is a human gene encoding a transcription factor broadly expressed in pancreatic progenitors and differentiated pancreatic cells. Proper regulation of *PDX1* expression is required in the initial stages of pancreatic organogenesis, as well for the determination of pancreatic cell lineages and maintenance of β-cell function [52].

Expression of *PDX1* in pancreatic progenitors is central to induct proper pancreatic organogenesis, which is demonstrated by pancreas agenesis in case of homozygous gene loss-of-function in humans. Moreover, several heterozygous missense and frameshift mutations on human PDX1 gene lead to distinct diabetic phenotypes,

highlighting PDX1's important role on adult pancreatic function. These diabetic phenotypes characterise MODY4, an early onset type of diabetes that comprehends different levels of hyperglycaemia, associated with dysregulation of glucose homeostasis that result from different degrees of *PDX1* functional impairment [96]. The complex regulation of *PDX1* is further demonstrated by phenotypes arising from mutations on regulatory elements. For instance, mice mutants display pancreatic agenesis upon deletion of conserved promoter regions [113]. Furthermore, cis-regulatory mutations and epigenetic alterations of *PDX1* are also associated to the development of T2D [114]. The distinct consequences of *PDX1* dysregulation highlight the worth of studying the transcriptional regulation of this *locus*. Therefore, the work presented in this thesis aims to uncover important cis-regulatory mechanisms that affect *PDX1* expression *in vivo* and might impact on diabetes development.

Zebrafish *pdx1* shows slightly different functions comparing to its human ortholog, as demonstrated in loss-of-function *pdx1* model. Homozygous *pdx1* null mutation in zebrafish leads to endocrine dysfunction and diabetic traits; on the other hand, distinctly from human and mouse, pancreatic organogenesis still occurs [87, 93]. Therefore, progenitor and adult human PDX1 CREs might be studied in the context of transgenesis reporter assays. Additionally, adult human PDX1 CREs might be studied functionally, since loss-of-function of *pdx1* does not generate a pancreatic agenesis phenotype. Overall, the PDX1 *loci* is the most adequate to reach the proposed objectives.

2. PDX1 *locus* analysis

1) Analysis of PDX1-linked SNP associated to T2D

The T2D-associated polymorphism rs35369009 (chr13:28,490,510, GRCh37/hg19/Human listed in **Table 3**) is positioned in an intron of a long non-coding RNA (lncRNA) gene, PDX1-associated lncRNA upregulator of transcription (*PLUT*). This antisense lncRNA was shown to promote interaction between the *PDX1* promoter and an upstream enhancer cluster, thus enhancing *PDX1* transcription [115]. Analysis of rs35369009 SNP reveals co-localization with an H3K27ac peak detected in adult pancreatic islets (**Figure 4**). Additionally, the SNP co-localizes with H3K4me1 peaks detected in human embryonic stem cell line (h1-HESC; data from the ENCODE Project). According to Parker S.C and colleagues, rs35369009 is contained within a stretch enhancer [28]. Stretch enhancers are defined as enhancer regions that are extended for a length of 3 kilobases (kb) or more, often laying on *locus* regulatory regions and showing

tissue-specific activity [28]. Interestingly, this SNP was also reported to be located within an islet active enhancer in the study of Pasquali and colleagues (**Figure 4;** [7]).

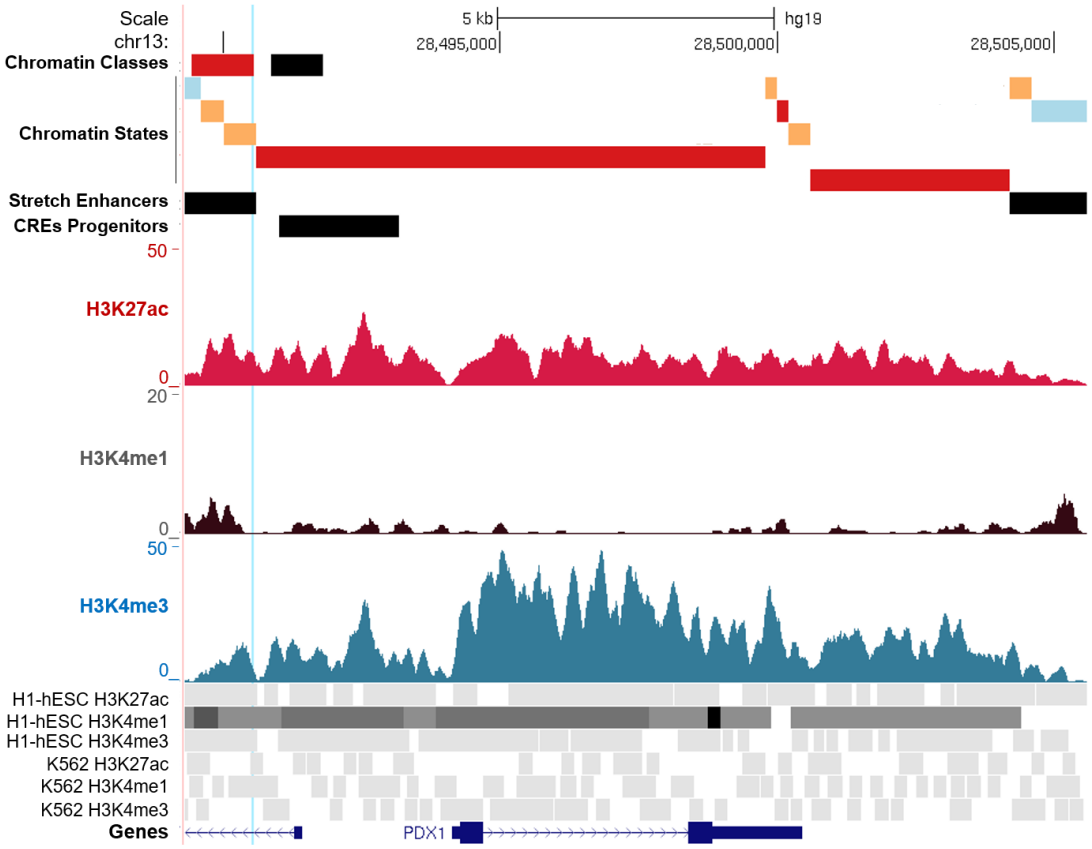

Figure 4. Epigenetic features characterising the T2D-associated rs35369009 SNP in the PDX1 *locus*. The top three tracks indicate functional prediction of the region adjacent to the SNP. Chromatin classes described by Pasquali L., and colleagues show annotations of enhancers (in red) and promoters (in black) detected in adult human islets [7]. Chromatin states defined by Miguel-Escalada I., and colleagues 2019 (Islet Regulome) include weak and strong enhancers (in light blue and orange, respectively) and active promoters (in red) [116]. The middle three indicate histones code analysis in human islets; the bottom tracks enhancer marks by ChIP-Seq in the correspondent cell lines. The SNP is vertically highlighted in cyan. The genomic coordinates are indicated at the top of the figure.

We also have investigated in detail the *in vivo* binding of the five key TFs operating in human islets [7] and additional TFs involved in early pancreas development in human pancreas multipotent progenitor cells (MPC; [39]). **Figure 5** (next) summarize this analysis, suggesting that there is no binding of PDX1 to its upstream region containing the SNP, while the other four key players could bind in that region, although presenting very low levels of binding. As there is no positive feedback by binding of PDX1 itself, it is worth to speculate that this region could be responsible for opening the *locus* at specific timepoints – in line with the onset of expression of the other factors, but not for sustaining

PDX1 transcription by a positive feedback loop mediated by PDX1 itself: *FOXA2* and *NKX2.2* are expressed earlier than *PDX1* during development, whereas *MAFB* and *NKX6.1* later. Thus, it could be possible that the activity of this putative regulatory element will be restricted to two distinct temporal stages.

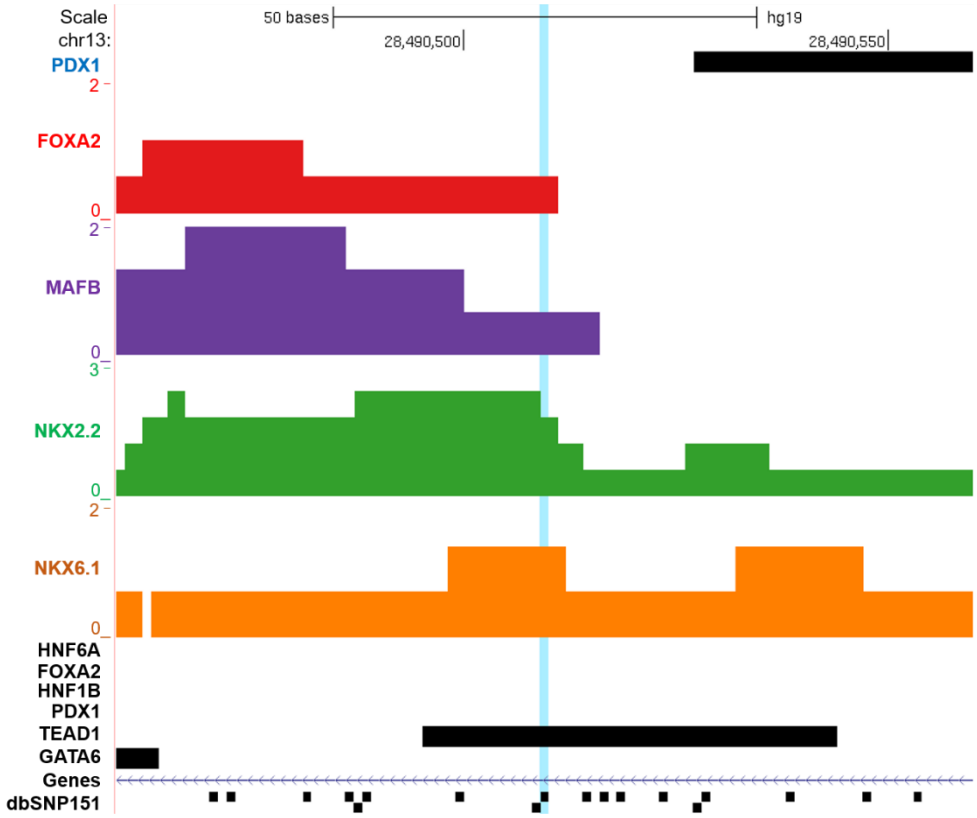

Figure 5. ChIP-Seq illustrates the *in vivo* binding of the indicated transcription factors in human islets (coloured tracks), as well as in progenitor cell lines (black tracks), around the T2D-associated rs35369009 SNP in the *PDX1 locus*. The SNP is vertically highlighted in cyan. The genomic coordinates are indicated at the top of the graph.

Moreover, TEAD1, known for regulating transcription of many genes in a context-specific-manner, was found to bind this region very mildly, in pancreatic progenitor cells, again suggesting an early role of this sequence during pancreatic development [39].

Overlapping of rs35369009 with epigenetic marks for enhancer activity and with binding sites of key-pancreatic TFs suggests that the association of the human polymorphism with diabetes could be explained as a result of the disruption of normal transcriptional cis-regulatory networks. Because PDX1 has a central role on pancreatic development and adult function, it will be crucial to investigate whether the disruption of *PDX1* regulatory mechanisms could be sufficient to cause a perturbation of glucose homeostasis and lead to other T2D-associated phenotypes.

2) Prediction of PDX1 CREs

In order to identify *PDX1* putative CREs, the *locus* was explored in the Islet Regulome Browser (GRCh37/hg19 /Human) (**Figure 6)**. Through the analysis of long-range interactions of the gene promoter sequence detected by pcHi-C, we were able to define putative limits of the *PDX1* regulatory landscape, being restricted between the coordinates chr13:28,397,000-28,510,000 (GRCh37/hg19 /Human). After defining the putative genomic coordinates of the PDX1 regulatory landscape, a screening for *PDX1* putative CREs was conducted. Sequences enriched in H3K27ac modifications detected in human pancreatic islets suggest the presence of three putative enhancers (hereafter referred as en1, en2 and eSNP). While the putative enhancer sequence termed eSNP is contained within a region showing enrichment for H3K27ac but not H3K4me1, the selection of the sequence of interest was mainly based on the location of the T2D-associated SNP, rs35369009. Sequences en1 and en2 were defined accordingly to overlap between regions showing enrichment in H3K27ac epigenetic mark and also interactions of the PDX1 promoter detected by pcHi-C. The genomic coordinates of the putative enhancers are presented in **Table 5.**

Table 5. Putative enhancer sequences selected in *PDX1 locus*.

Sequence	Genomic coordinates (GRCh37/hg19)	Putative enhancer length (bp)
en1 PDX1	chr13:28413883-28415558	1676
en2 PDX1	chr13:28442455-28443986	1532
eSNP PDX1	chr13:28489417-28490575	1159

Furthermore, we predicted the presence of insulators in the PDX1 regulatory landscape, using CTCF and cohesin ChIP-Seq tracks, as these proteins are associated to insulator functions, establishing barriers within adjacent genes landscapes, as well as modulating enhancer activity [20]. Three sequences were selected, two upstream the *PDX1* coding sequence (named ins1 and ins2), and one region downstream (ins3; **Figure 6**). It will be interesting to understand whether these are the regions defining the TAD of PDX1, or whether they are acting within a bigger TAD to modulate *PDX1* transcription by modulating the activity of other enhancers. The genomic coordinates of the putative insulators are presented in **Table 6**.

Table 6. Putative insulator sequences selected in *PDX1 locus*.

Sequence	Genomic coordinates (GRCh37/hg19)	Putative insulator length (bp)
ins1 PDX1	chr13:28400284-28401759	1476
ins2 PDX1	chr13:28403943-28404737	795
ins3 PDX1	chr13:28503601-28504743	1143

The putative CREs, both insulators and enhancers, overlap with open regions of chromatin detected by ATAC-Seq in human islets (**Figure 6**). The availability of these sequences is consistent with the requirement for those to be accessible to the binding of TFs.

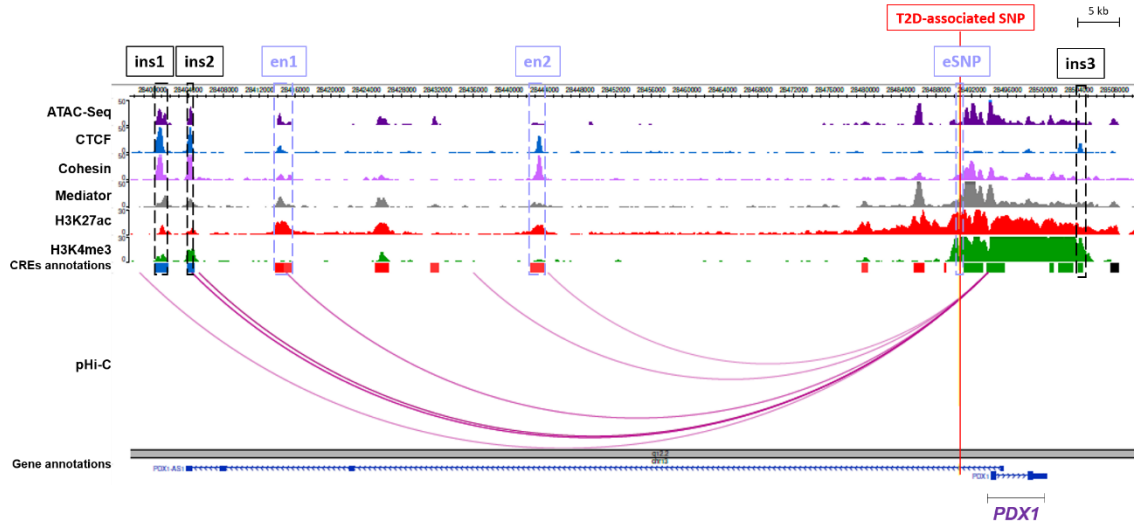

Figure 6. Screening for putative CREs and defining the *PDX1* landscape. On top of the picture, the scale is indicated; below, dashed rectangles highlight the selected putative CREs with respective names. The first track corresponds to ATAC-Seq data showing regions of open chromatin; the following three tracks shows genomic regions enriched in CTCF, cohesin and Mediator binding detected by ChIP-Seq; next two tracks show histone enrichment in H3K4me1 and H3K27ac epigenetic marks that define enhancer regions; below, there are presented annotations of putative regulatory elements; and finally, the last track shows chromatin interactions detected by pcHi-C, with a viewpoint in PDX1 promoter. All the data presented in the figure corresponds to assays performed in human pancreatic islets. At the bottom of the figure, genes are illustrated.

3. Validation of PDX1 CREs in zebrafish

1) Subcloning of putative CREs sequences into entry vector

The sequences selected as *PDX1* putative CREs were amplified by PCR. To do so, primers flanking these sequences were designed, which are listed in **Table 6** (see Materials and Methods, **Table 3**). PCR amplification was performed using as template

DNA, a PDX1 BAC containing the human genomic regions of interest or a sample of human genomic DNA (gDNA) (**Figure 7 A** and **B,** respectively). PCR reactions were run in a 1% (w/v) agarose gel and amplicons were confirmed to be the expected molecular weight (**Figure 7** and **8**).

Putative CREs sequences were amplified first from the PDX1 BAC and then from gDNA from healthy patients for technical and biological reasons. First, it is easier to optimize a PCR employing single molecule as a BAC than a higher-complexity DNA sample as genomic DNA. More importantly, as the final aim of this thesis is to introduce

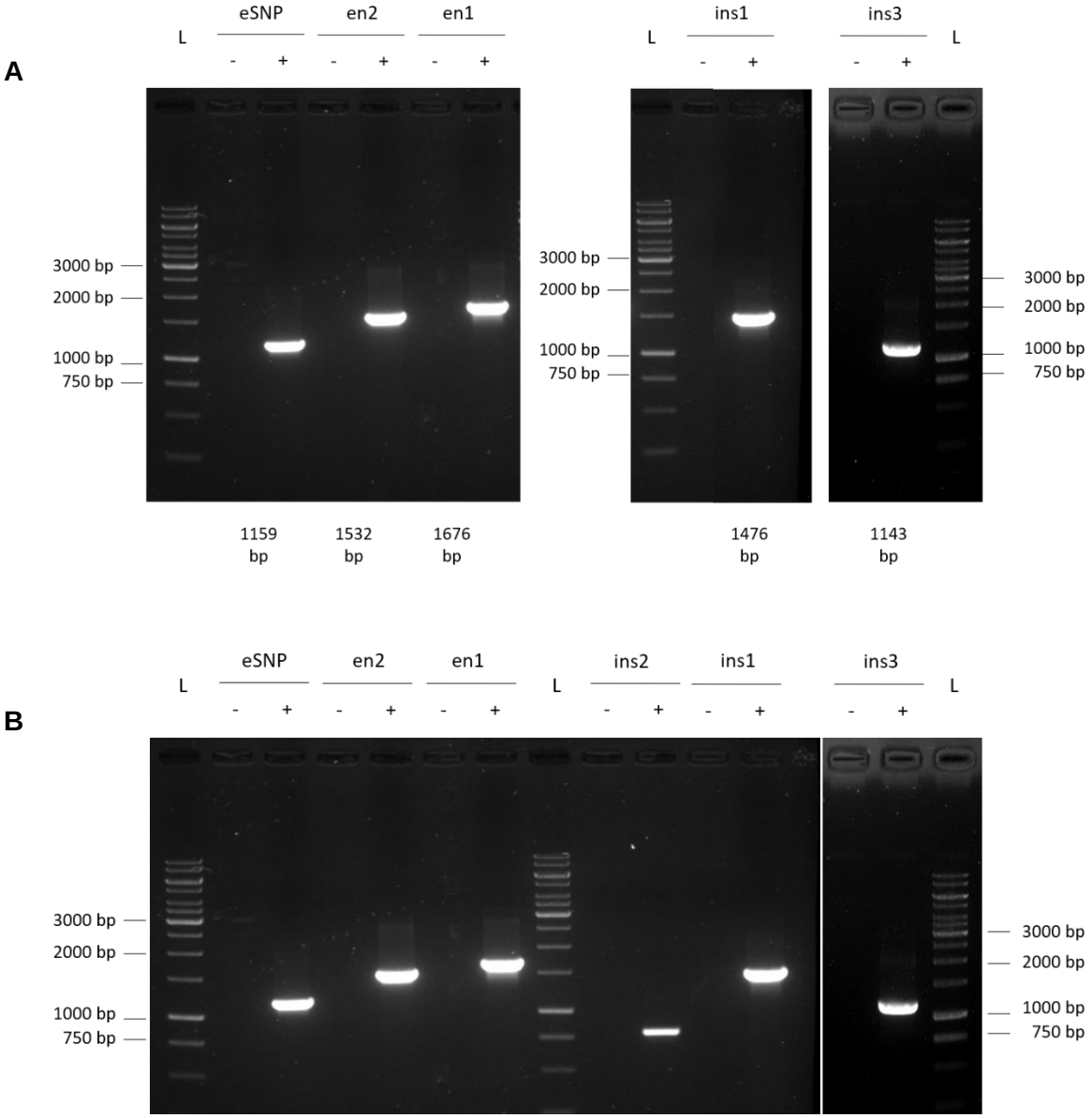

Figure 7. Electrophoresis gel of PCR amplification of PDX1 putative CREs from (**A**) BAC DNA template and (**B**) gDNA template. Symbols "-" and "+" refer to PCR negative control (blank) and DNA-containing PCR mixes, respectively; while “L” represents the 1 kb ladder. The amplified regions are indicated on the top of the gel, while their correspondent size is depicted at the bottom.

a human PDX1 BAC into the zebrafish genome (see Chapter 4), it is required to evaluate the exact sequences situated in the PDX1 BAC, as they might carry mutations.

Interestingly, the ins2 region was found to be amplified exclusively when using template from human gDNA (**Figure 7**), failing when using BAC DNA, even after optimization through a gradient of annealing temperatures between 54 and 60 °C (**Figure 8**). Absence of PCR product derived from the BAC could be better explained by explained by the presence of a mutation, namely in the region where the primer pair was designed to anneal, thus impairing amplification of this putative CRE, or alternatively by a deletion. New primers should be designed spanning a larger sequence, flanking the selected region. Another possible explanation for the absence of PCR product when using the PDX1 BAC as template could be caused by the formation of DNA secondary structure due the sequence itself, which might be more stable in a single molecule as the BAC template, although this hypothesis is less likable since gDNA PCR product is obtained.

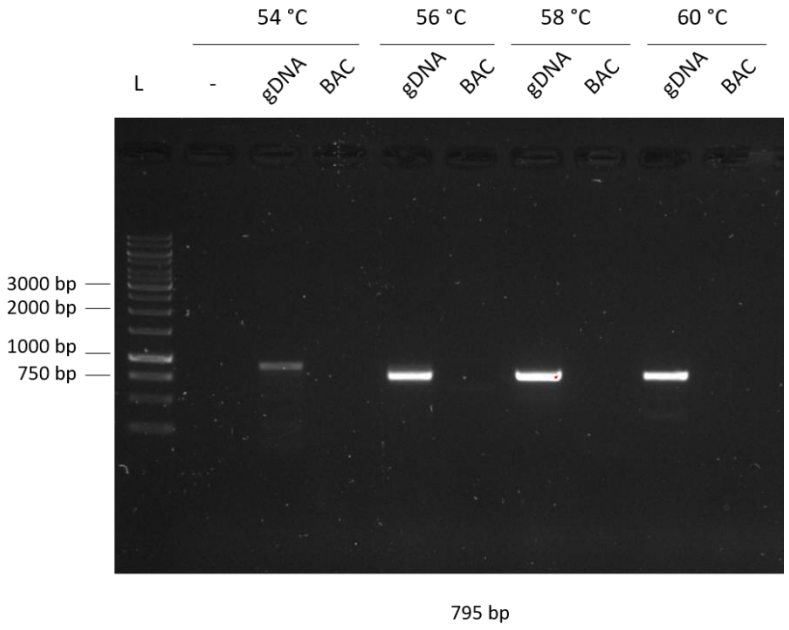

Figure 8. Electrophoresis gel of PCR amplification of PDX1 putative CRE ins2 from gDNA and BAC templates. Symbols "-" and "+" refer to PCR negative control (blank) and DNA-containing PCR mixes, respectively; while “L” represents the 1 kb ladder. The annealing temperatures and the templates are indicated on the top of the gel, while the correspondent size is depicted at the bottom.

After amplification of the PCR products, the amplicons were cloned into the entry vector pCR™8/GW/TOPO®. Success of cloning reactions was then confirmed by

restriction of extracted plasmids with *EcoRI* restriction enzyme. pCR™8/GW/TOPO® vector includes two *EcoRI* restriction sites flanking the gateway site (see Materials and Methods). Consequently, digestion of the vector containing the putative CREs cloned within should result in two linear DNA fragments, one with a size of 2817 bp (vector backbone) and the other with the size of the correspondent sequence contained. Digestion products were analysed in a 1% agarose gel and confirmed the successful cloning of each putative CRE into the entry vector (**Figure 9**).

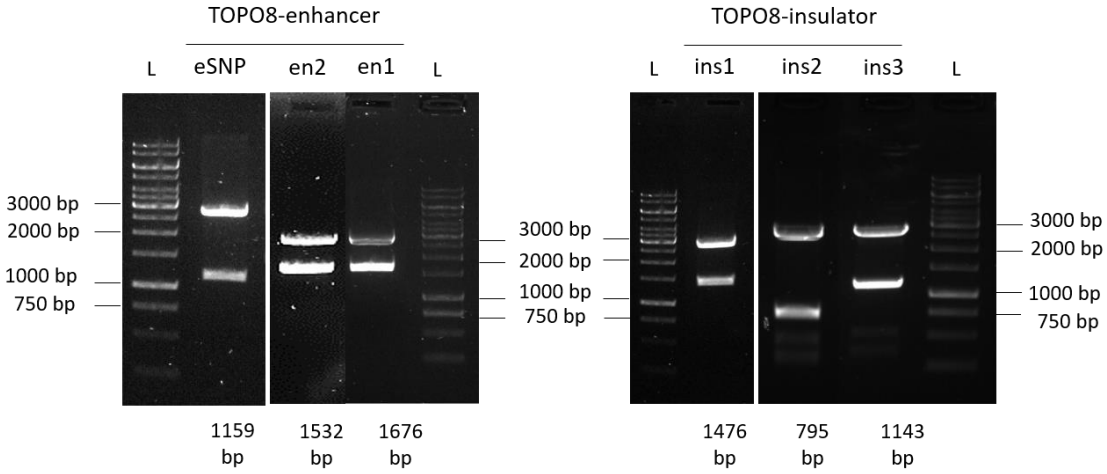

Figure 9. Representative electrophoresis gel of pCR™8/GW/TOPO® vector containing PDX1 putative CREs after digestion with *EcoRI* restriction enzyme. Each lane of the gel shows a 2817 bp band (vector backbone) and the band corresponding to the size of the cloned sequence. Symbol "L" represents the 1 kb ladder.

(1) CREs sequences analysis derived from BAC and gDNA PCR amplification

For each PDX1 putative CRE cloned into the entry vector pCR™8/GW/TOPO®, the presence of the correct sequence in the entry vector was further confirmed by Sanger-sequencing. Sequencing results were analysed by alignment of reads with the most recent version of the human DNA available in UCSC Browser (GRCh37/hg19; **Tables 4** and **5**). Sequencing results allowed the confirmation of correct cloning of *PDX1* putative CREs in the entry vector and, importantly, alignment allowed to detect human variants in the cloned sequences, that were later tested as CREs.

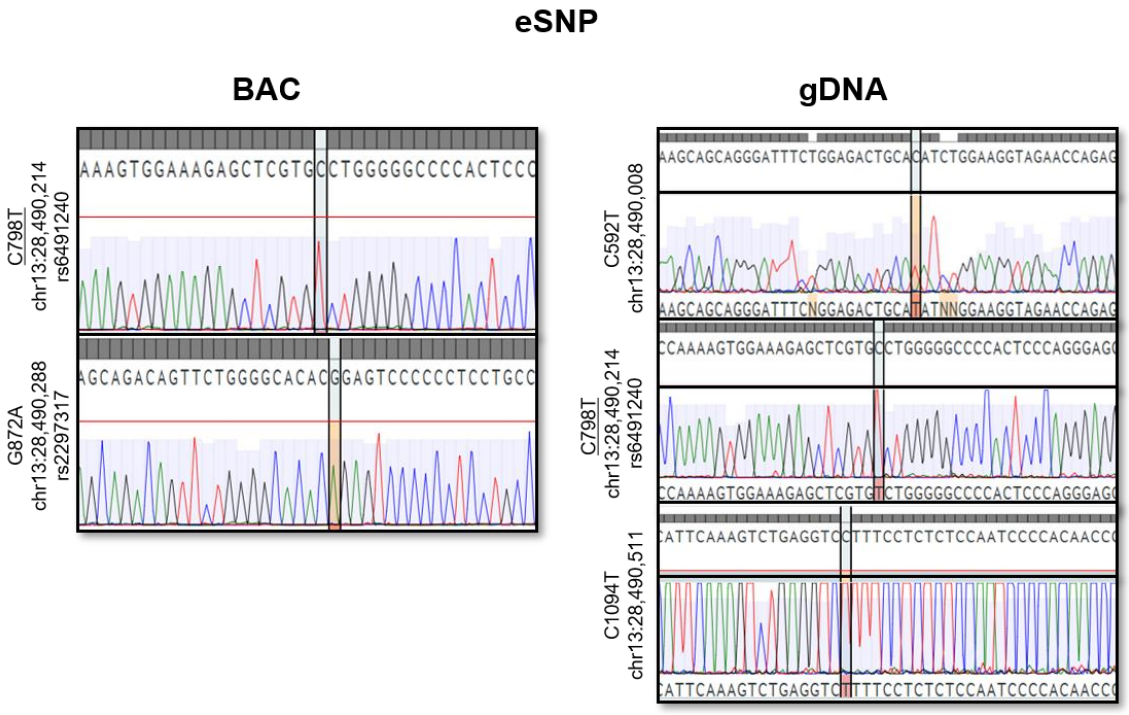

Figure 10. Alignment of sequencing reads of eSNP cloned in pCR™8/GW/TOPO® with the human genome (GRCh37/hg19). Each box depicts regions where polymorphisms were detected; at the side genomic coordinates and the mutations are illustrated. When the mutations are described as common SNPs are shown together with the correspondent code (format 'rs#'). Polymorphisms detected in sequences amplified both from the BAC DNA and gDNA templates are presented underlined.

The eSNP was amplified from the BAC and human gDNA. The gDNA sequencing results revealed three single nucleotide variants (SNVs) comparing to the reference sequence. One is described as a common polymorphism (that is, a SNP) (rs6491240), while the remaining two variants located at chr13:28,490,008 and chr13:28,490,511 were not annotated in dbSNP database [117]. Furthermore, the sequence obtained from the BAC contains two SNPs, with the following genomic coordinates: chr13:28,490,214 and chr13:28,490,288, which were both previously annotated in the dbSNP database as rs6491240 and rs2297317, respectively (**Figure10**) [117]. More importantly, the T2D-associated SNP located in this region was not detected on the sequencing results obtained, both from BAC DNA and gDNA samples, being the ideal BAC clone to use for humanizing the zebrafish genome, as proposed in Chapter 4.

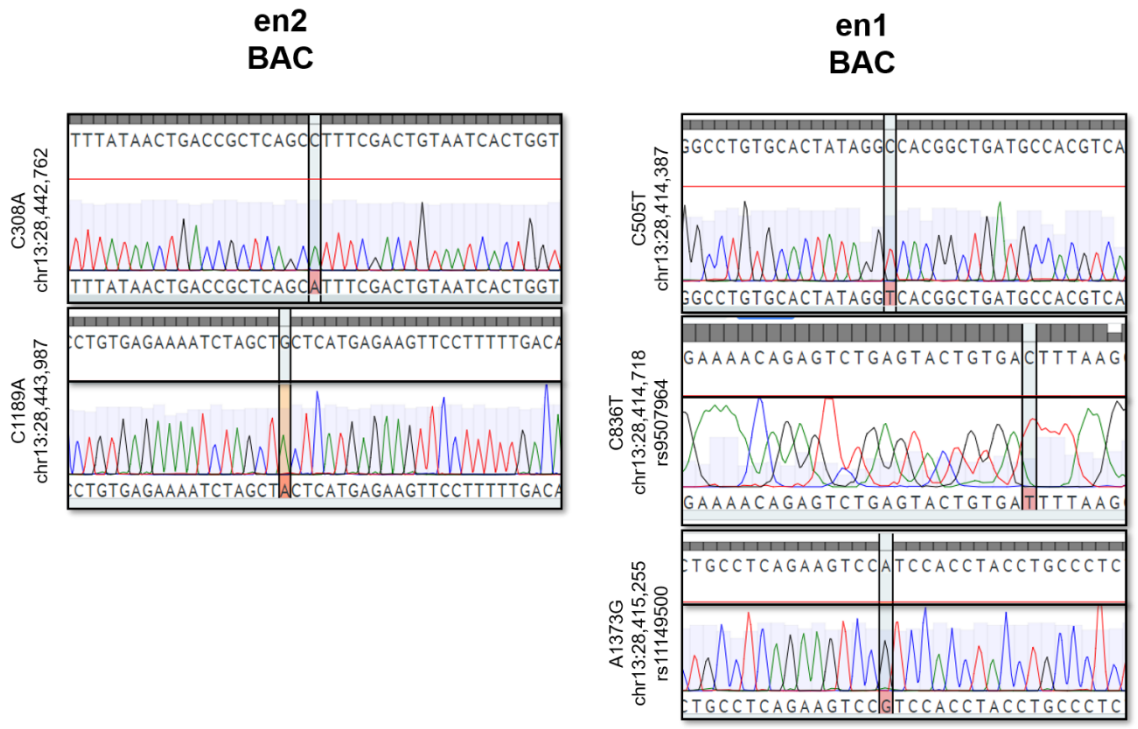

Figure 11. Alignment of sequencing reads of en2 and en1 cloned in pCR™8/GW/TOPO® with the human genome (GRCh37/hg19). Each box depicts regions where polymorphisms were detected; at the side genomic coordinates and the mutations are illustrated. When the mutations are described as common SNPs are shown together with the correspondent code (format 'rs#').

Alignment of reads from pCR™8/GW/TOPO® containing the putative enhancers en1 and en2, amplified using the PDX1 BAC as template DNA, confirmed the successful cloning of the desired sequences. In the case of en1, the amplified sequence from the BAC DNA led to identify three SNVs, two of them annotated in dbSNP database (rs9507964 and rs11149500; **Figure 11**). An extra variant, not annotated in the annotated in dbSNP database, was detected in the position chr13:28,414,387 corresponds to a C>T nucleotide modification. Regarding the en2 sequence, two SNVs were identified, none of them described in dbSNP database (**Figure11**).

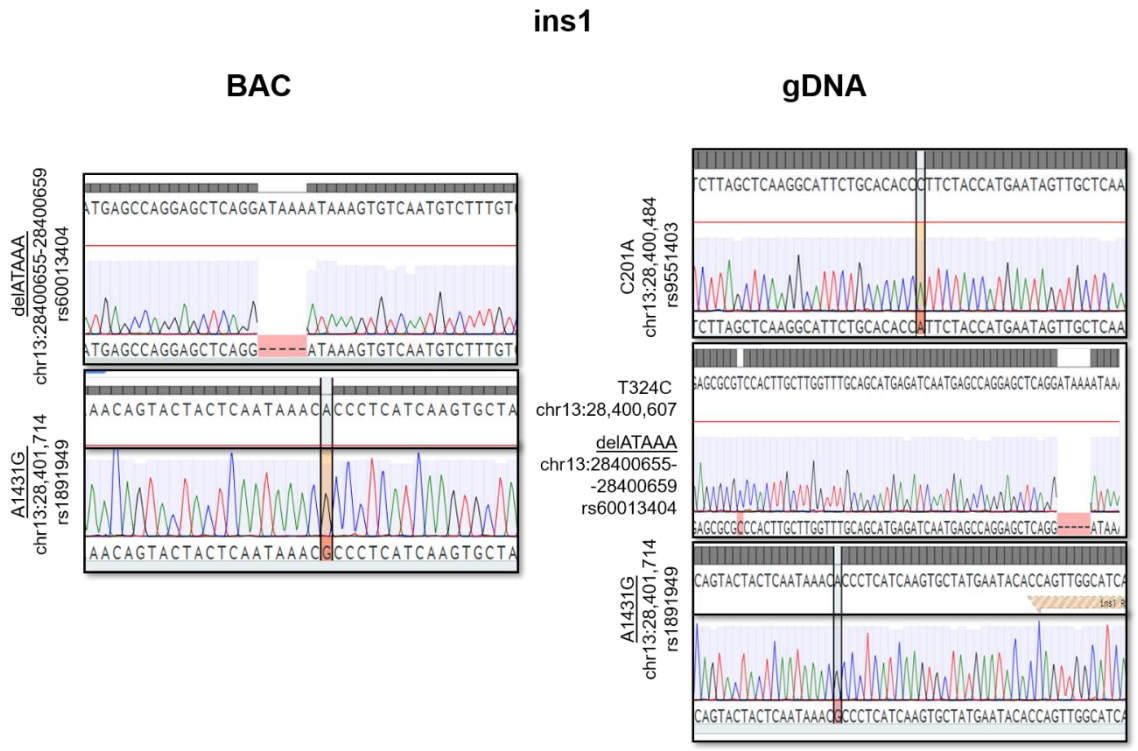

Figure 12. Alignment of sequencing reads of ins1 cloned in pCR™8/GW/TOPO® with the human genome (GRCh37/hg19). Each box depicts regions where polymorphisms were detected; at the side genomic coordinates and the mutations are illustrated. When the mutations are described as common SNPs are shown together with the correspondent code (format 'rs#'). Polymorphisms detected in sequences amplified both from the BAC DNA and gDNA templates are presented underlined.

The sequence ins1 was amplified both from BAC DNA and human gDNA. On the sequence amplified from BAC DNA, two variants were detected: the first corresponds to a five-nucleotides deletion (ATAAA) and the second to a SNP (**Figure 12**). Both the ATAAA deletion and the single nucleotide transition A>G have been described in dbSNP database. Regarding human gDNA, ins1 sequence contains four SNVs comparing to the reference genome: three of those were already reported in dbSNP, while the nucleotide transition at position chr13:28,400,607 was not reported (**Figure 12**).

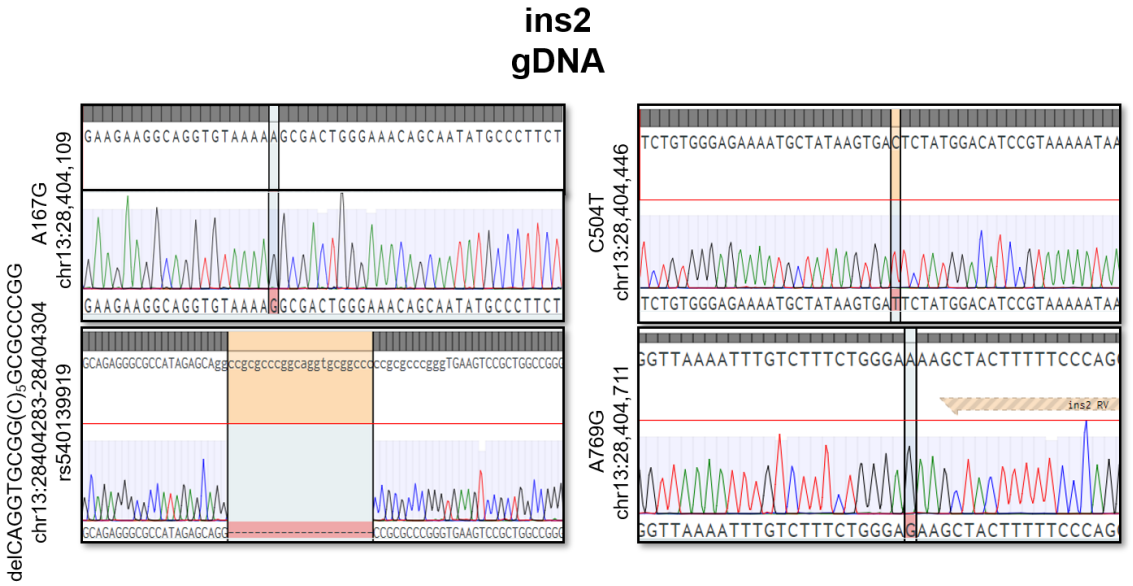

Figure 13. Alignment of sequencing reads of en2 and en1 cloned in pCR™8/GW/TOPO® with the human genome (GRCh37/hg19). Each box depicts regions where polymorphisms were detected; at the side genomic coordinates and the mutations are illustrated. When the mutations are described as common SNPs are shown together with the correspondent code (format 'rs#').

Alignment of reads from pCR™8/GW/TOPO® containing the putative insulator ins2 amplified from gDNA allowed to detect four variants: three SNVs and one deletion of 22-nucleotides. None of the SNVs is described in dbSNP database (**Figure 13**). Interestingly, the deletion of 22 nucleotides within the putative insulator has already been described in dbSNP. Moreover, the detection of this deletion along with the other three SNPs reinforces the hypothesis that amplification of ins2 from BAC DNA could indeed have been impaired due to presence of mutations in this template sequence.

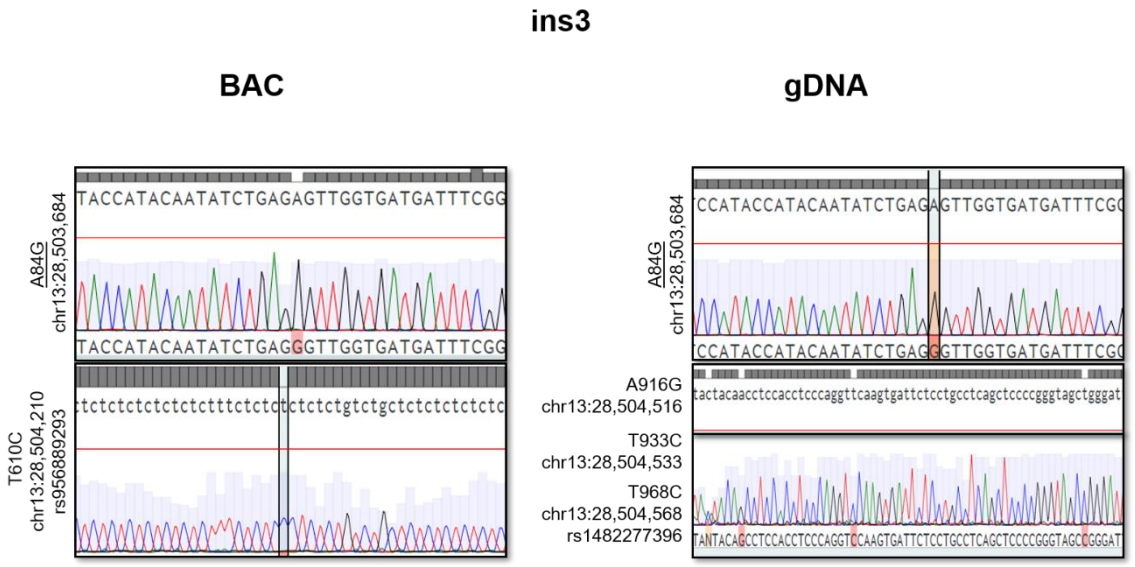

Figure 14. Alignment of sequencing reads of ins3 cloned in pCR™8/GW/TOPO® with the human genome (GRCh37/hg19). Each box depicts regions where polymorphisms were detected; at the side genomic coordinates and the mutations are illustrated. When the mutations are described as common SNPs are shown together with the correspondent code (format 'rs#'). Polymorphisms detected in sequences amplified both from the BAC DNA and gDNA templates are presented underlined.

Finaly, sequencing results obtained from the pCR™8/GW/TOPO® construct containing the putative insulator ins3 confirmed successful cloning of sequences amplified both from BAC DNA and gDNA. Alignment of reads with the reference genome led to detect two SNVs on the BAC ins3 sequence (**Figure 14**). Of these two, only one was reported in dbSNP database (rs956889293). In case of gDNA ins3 sequence, four SNVs were detected. Of those, only the most downstream variant is reported in dbSNP. Remarkably, despite unreported in the database, the variant detected at position chr13:28,503,684 was found to be included in ins3 sequences amplified from both DNAs.

A summary of these results is presented in **Table 7**. Considering the size of each of the putative CREs (from 700 bp to 2 kb, approximately), overall no drastic sequence aberrations were detected. However, many variants are present in these sequences, some being already annotated in the SNP database as common SNPs, but some are not (referred as SNVs). A cautionary note must be highlighted regarding especially the second class of variants, since they are not described as common SNPs, opening the possibility of having deleterious consequences in CREs activity. Nevertheless, considering that most annotated SNPs derive from exome-sequencing, it is likely that the statistical significance in non-coding regions to define common SNPs are not so powerful yet. Moreover, there are many SNPs/mutations common between the two human *PDX1* sequences (BAC and gDNA; summarized in **Table 7**), indicating that these variants can

indeed be more prevalent, although not annotated. Lastly, this characterisation at single nucleotide resolution will be very important when combining with *in vivo* CREs reporter assays, to determine if some of these variants might have an impact be the CRE activity.

Table 7. Summary of mutations in the amplified indicated sequences, classified for the template used, either the PDX1-BAC or gDNA. For annotated SNPs, their reference is reported. Common mutations between the two templates are underlined. *Del stands for delCAGGTGCGG(C)5GCGCCCGG.

Putative CREs	BAC			gDNA		
	Polymorphism	Genomic coordinates (GRCh37/hg19)	dbSNP code	Polymorphism	Genomic coordinates (GRCh37/hg19)	dbSNP code
eSNP				C592T	chr13:28,490,008	–
	C798T	chr13:28,490,214	rs6491240	C798T	chr13:28,490,214	rs6491240
	G872A	chr13:28,490,288	rs2297317	C1094T	chr13:28,490,511	–
en2	C308A	chr13:28,442,762	–			
	C1189A	chr13:28,443,987	–			
en1	C505T	chr13:28,414,387	–			
	C836T	chr13:28,414,718	rs9507964			
	A1373G	chr13:28,415,255	rs11149500			
ins1				C201A	chr13:28,400,484	rs9551403
				T324C	chr13:28,400,607	–
	delATAAA	chr13:28400655-28400659	rs60013404	delATAAA	chr13:28400655-28400659	rs60013404
	A1431G	chr13:28,401,714	rs1891949	A1431G	chr13:28,401,714	rs1891949
ins2				A167G	chr13:28,404,109	–
				Del*	chr13:28,404,283-28,404,304	rs540139919
				C504T	chr13:28,404,446	–
				A769G	chr13:28,404,711	–
ins3	A84G	chr13:28,503,684	–	A84G	chr13:28,503,684	–
	T610C	chr13:28,504,210	rs956889293	A916G	chr13:28,504,516	–
				T933C	chr13:28,504,533	–
				T968C	chr13:28,504,568	rs1482277396

2) Recombination of putative CREs sequences into destination vector

The sequences cloned pCR™8/GW/TOPO® were further transferred into destination vectors using gateway recombination, which were then used to functionally evaluate the role of these human sequences as CREs. The destination vectors are: (1) the Z48 vector for putative enhancers and (2) the insulator test vector for putative insulators.

(1) Enhancer test vector

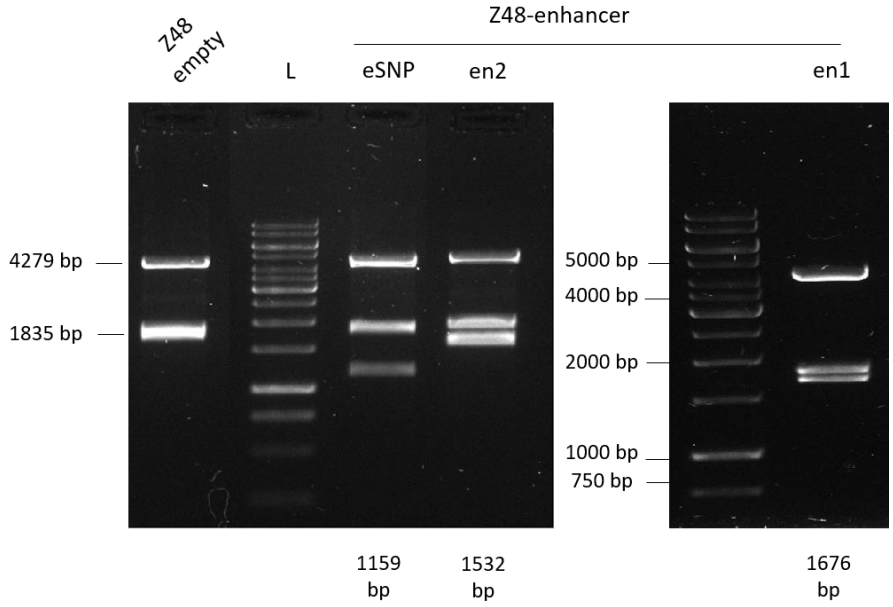

Figure 15. Representative electrophoresis gel of Z48 vector containing PDX1 putative CREs after digestion with *EcoRI* restriction enzyme. Symbol "L" represents the 1 kb ladder. The first lane shows the digestion pattern of the Z48 empty vector. The other lanes illustrate specific digestion pattern of successfully cloned regions into Z48.

Putative enhancers were recombined into the Z48 transposable element. Successful recombination of putative enhancer sequences was tested by digestion of extracted plasmids with *EcoRI* restriction enzyme. Similarly to pCR™8/GW/TOPO®, Z48 vector contains two *EcoRI* restriction sites flanking the cloning site. Additionally, the vector contains one more restriction site placed in-between the GFP cassette and Z48 enhancer. Consequently, digestion of the vector containing the putative CREs cloned within should result in three DNA fragments, two with sizes of 4279 bp and 1835 bp (vector backbone) and the third with the size of the cloned putative CRE. Digestion

products were analysed in a 1% agarose gel and confirmed the successful cloning of each putative enhancer into Z48 vector (**Figure 15**).

(2) Insulator test vector

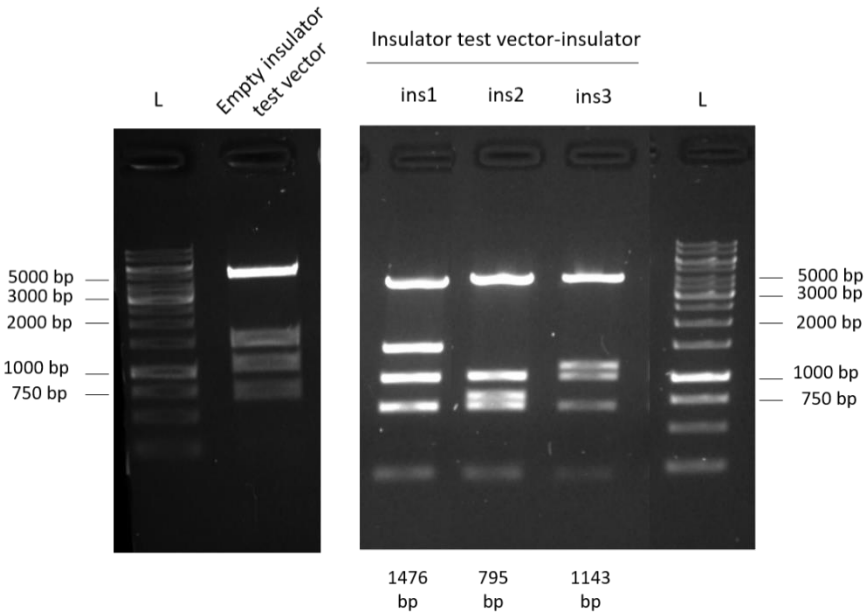

Figure 16. Representative electrophoresis gel of insulator test vector containing PDX1 putative CREs after digestion with *EcoRI* restriction enzyme. Symbol "L" represents the 1 kb ladder. The second lane shows the empty vector after digestion, resulting in four bands. The other lanes illustrate specific digestion pattern of successfully cloned regions into insulator test vector.

Putative insulators were recombined into the insulator test vector and successful recombination of putative insulator sequences was tested by *EcoRI* digestion of plasmids. The insulator test vector contains four *EcoRI* recognition sequences, two of them flanking the vector cloning site. Restriction of the empty vector (without a cloned sequence) results in multiple bands, presented in the first lane of the electrophoresis gel (**Figure 16**) – specific size of bands is not shown since the full sequence of insulator test vector was not available. Alternatively, the successful cloning was confirmed by difference in the digestion pattern from the empty vector. Additionally, the size of the bands obtained upon digestion was approximately calculated from the gel and their resulting sum confirmed the insertion of the expected sequences (**Figure 16**).

3) Transgenesis assays to test CREs activity

(1) Test of PDX1 putative enhancers

In order to test the function of the human sequences selected as pancreatic enhancers, each Z48 vector containing putative *PDX1* enhancers was injected in one-cell stage zebrafish embryos along with *Tol2* mRNA [106]. Because *PDX1* is expressed in endocrine differentiated cells and pancreas progenitor cells [52], we searched for ways to label these different zebrafish cell types. To help locate the zebrafish pancreas, we took advantage of a *Somatostatin*-mCherry (*sst*-mCherry) zebrafish reporter line, where the *sst2* promoter drives the expression of the *mCherry* reporter gene in δ-cells since 17 hpf [62], therefore consistently indicating the endocrine pancreas domain. Moreover, and because *PDX1* is also expressed in the pancreatic progenitor domain, we labelled pancreatic progenitor cells using an anti-Nkx6.1 antibody [118]. In this experiment, we choose the 36 hpf time point since it is possible to detect differentiated endocrine cells and pancreas progenitor cells. A positive *PDX1* CRE in this pancreatic enhancer reporter assay is therefore expected to drive GFP expression either within the *sst*-mCherry domain or the Nkx6.1 pancreatic progenitor domain.

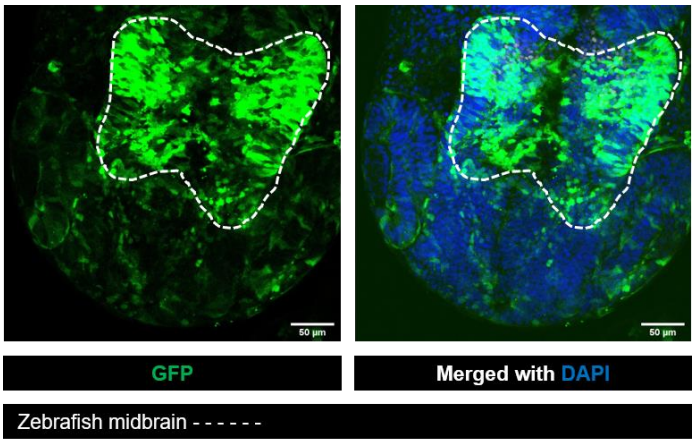

Figure 17. Representative confocal image of GFP expression in midbrain of a zebrafish embryo efficiently microinjected with a Z48 transposable element along with *Tol2* mRNA. The dashed white line represents zebrafish midbrain. Leica confocal SP5II; zoom 1x; magnification 40x.

After microinjection of the Z48 vector, the Z48 enhancer triggers GFP expression in the midbrain of 24hpf zebrafish embryos, as illustrated in **Figure 17**, serving as an internal positive control for transgenesis (**Figure 18**). After selection of embryos that show expression of GFP in the midbrain, expression of GFP in the differentiated or progenitor domains is then assayed, using confocal microscopy and the abovementioned reporters.

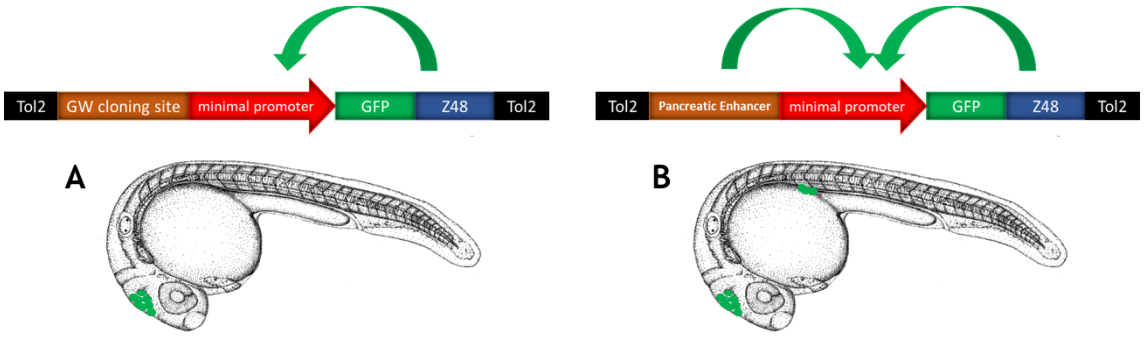

Figure 18. Scheme of the Z48 vector and GFP expression driven in zebrafish. (A) Empty Z48 vector. GFP expression is driven in the midbrain due to interaction of the Z48 midbrain enhancer with the minimal promoter upstream the reporter gene. (B) Z48 vector containing a cloned pancreatic enhancer upstream the minimal promoter. GFP expression is driven in zebrafish midbrain and pancreas, due to interaction of the Z48 and the pancreatic enhancer, respectively, with the minimal promoter.

In this assay, the Z48 vector without a recombined CRE (empty Z48) was injected as a negative control. Out of fifteen embryos microinjected with the empty Z48 vector, one showed GFP expression in endocrine domain (**Figure 19 A and B**). The GFP expression detected in the pancreas of this embryo might be explained by a phenomenon called “position effect” [15]. *Tol2* transposase integrations occur randomly in the genome, with some of those being nearby pancreatic enhancers, having the potential to induce the expression of GFP in pancreatic cells. This negative control anticipates that the threshold of noise for this assay is one out of fifteen (6,7%) GFP positive embryos. However, the number of analysed embryos is yet small and should be increased in future experiments.

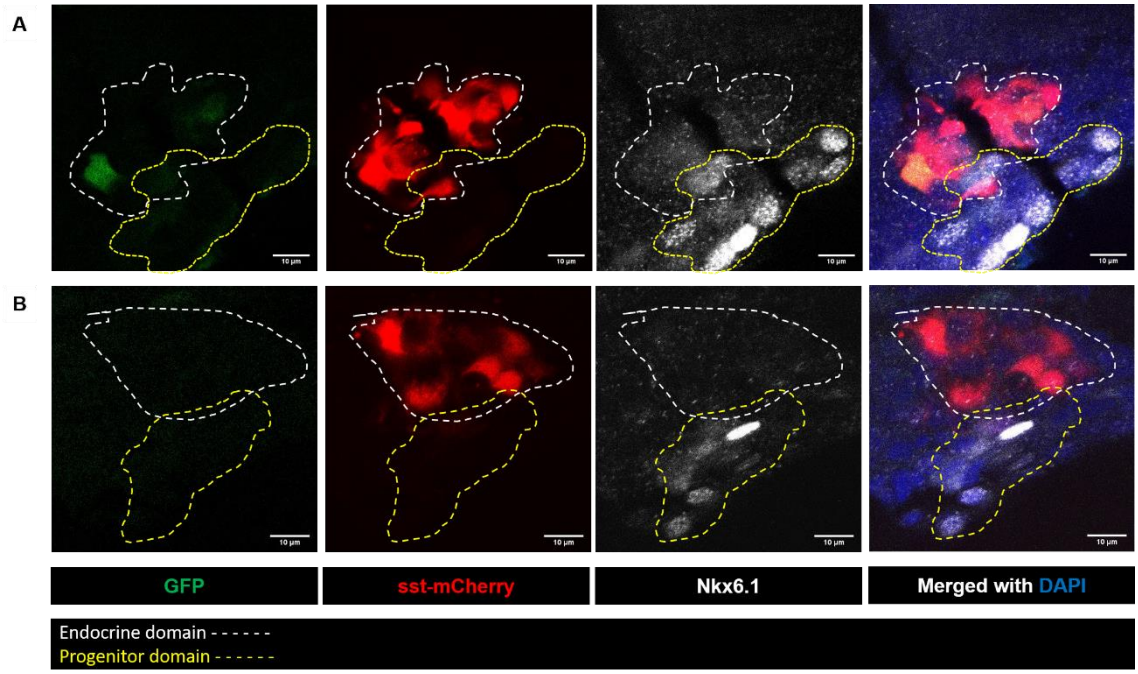

Figure 19. Representative confocal images of the pancreatic domain of embryos at 36 hpf injected with empty Z48 vector. Red fluorescence represents sst-expressing cells from the transgenic line tg(sst:mCherry). (A) Representative image of GFP expression in the endocrine domain. (B) Representative image of absence of GFP expression in the pancreatic domain. Leica confocal SP5II; zoom 2,91x; magnification 40x.

To determine the sensibility of this assay to detect endocrine enhancers, a sequence previously validated as a pancreatic enhancer through luciferase reporter assays performed in a mice β-cell line, was used as positive control [7]. As preliminary data, out of eight embryos analysed, two showed GFP expression in the endocrine domain (25%; **Figure 20**). The number of analysed embryos is yet small and should be increased in future experiments.

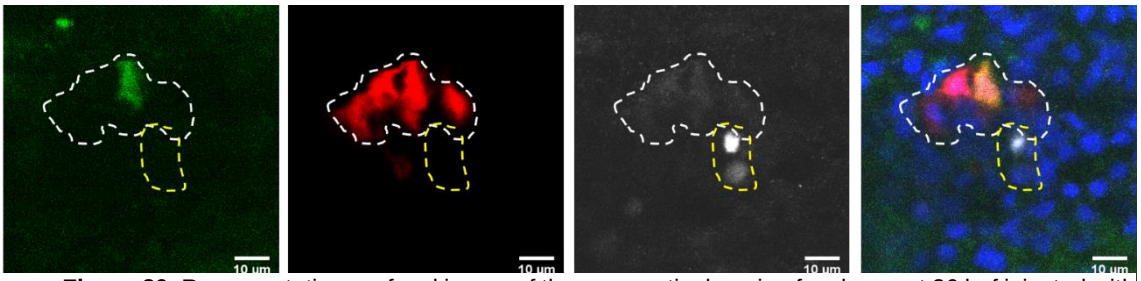

Figure 20. Representative confocal image of the pancreatic domain of embryos at 36 hpf injected with the Z48 vector containing a previously established enhancer. Red fluorescence represents the endocrine domain (sst-expressing cells), which was found to co-localize with GFP expression. Leica confocal SP5II; zoom 2,91x; magnification 40x.

Regarding the PDX1 putative enhancers, three *loci* were tested (eSNP, en1 and en2; **Figure 6**). For the eSNP *locus*, two sequences were amplified, one from the respective BAC DNA (eSNP BAC; BAC clone CH17-423D7) and another from genomic DNA (eSNP gDNA). For sequence eSNP gDNA, one out of seventeen embryos (5,9%) showed GFP expression in the pancreatic progenitor domain, and none in the pancreatic endocrine domain (**Figure 21**). For sequences en1 BAC (N=13) and en2 BAC (N=12), none of the analysed embryos showed GFP expression in the pancreatic endocrine or progenitor domain. Representative confocal images referring to the analysis of each sequence are shown in **Figure 21**.

Overall, the data provided through this assay was not able to validate or exclude the activity of the three analysed *loci* as enhancers, from the *PDX1* genomic landscape. The positive control showed a predicted tendency for having an increased number of embryos with GFP expression in the pancreatic endocrine domain, when comparing with the negative control (25% vs 6,7%). However, because the number of analysed embryos is yet limited, this difference is not statistically significant (P-value > 0,05 in unpaired t-test with Welch's correction; P-value = 0,3255). In future experiments an increased number of embryos must be analysed. Additionally, it is possible that the absence of GFP expression detected in the pancreatic domain of the majority of the zebrafish embryos analysed in different experimental conditions outcomes from low transgenesis efficiency. An inefficient integration of the Z48 transposable element in the zebrafish genome might lead to a decrease in the expression of the GFP reporter contained within, therefore being insufficient to detect that expression in the pancreatic domain. This can be improved by using quantitative assays to determine expression of GFP in the midbrain to evaluate the efficiency of transgenesis, as an alternative to the qualitative appreciation currently performed. Finally, we could also hypothesize that the sequences in test are unable to enhance GFP expression in the chosen timepoint. That is, *PDX1* enhancers might not be active at 36 hpf in zebrafish, becoming active at a later developmental stage, in mature pancreatic endocrine cells.

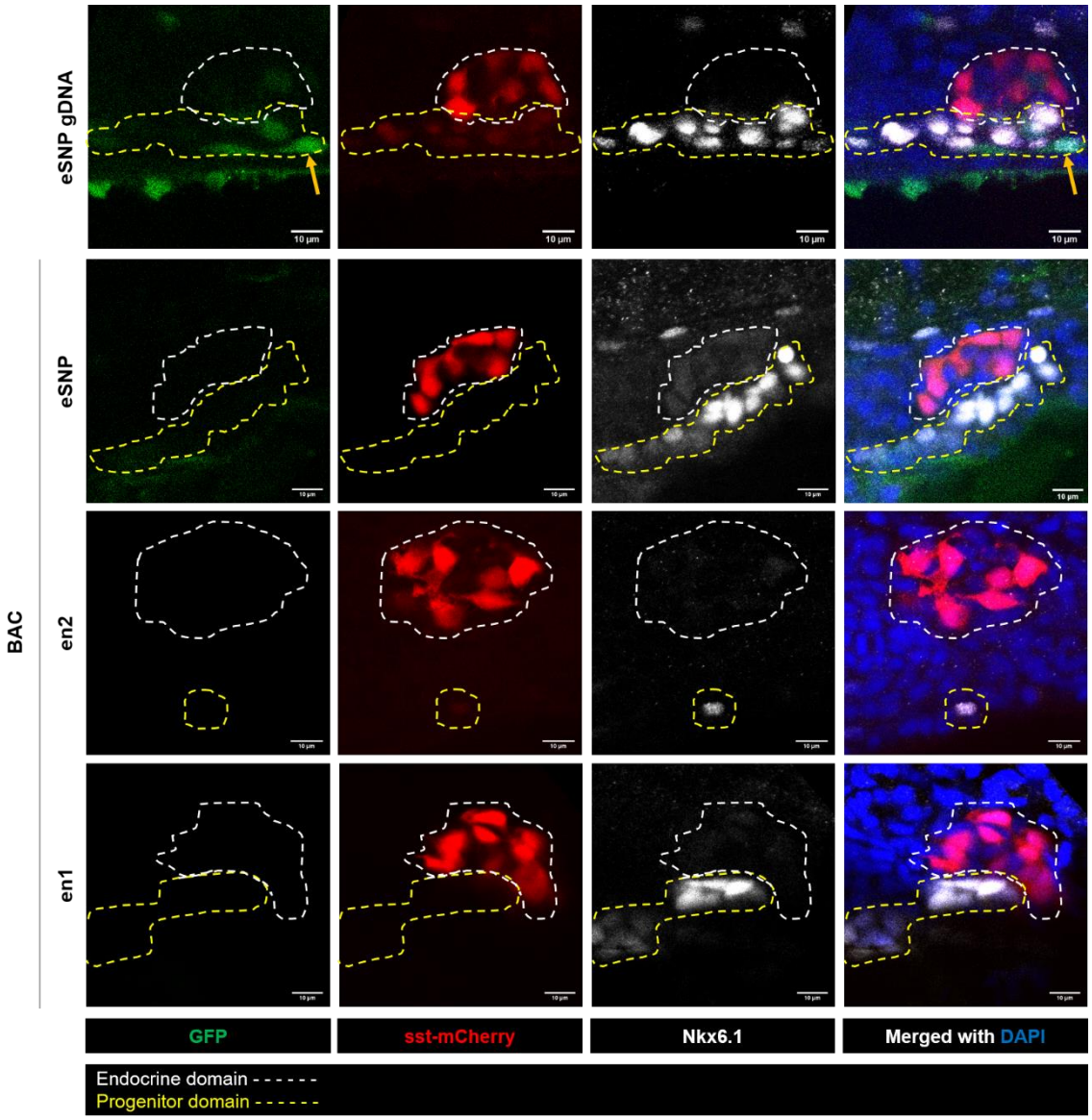

Figure 21. Representative confocal images of the pancreatic domain of embryos at 36 hpf injected with: eSNP amplified from BAC DNA (eSNP BAC) and from human gDNA (eSNP gDNA); en2 and en1 amplified from BAC DNA (en2 BAC and en1 BAC). Red fluorescence represents sst-expressing cells from the transgenic line tg(sst:mCherry). Yellow arrows are pointed to GFP-positive cells. Leica confocal SP5II; zoom 2,91x; magnification 40x.

(2) Test of PDX1 putative insulators

In order to test whether the human sequences selected and shown in **Table 5** can function as insulators, each insulator test vector containing putative *PDX1* insulators was injected in one-cell stage zebrafish embryos along with *Tol2* mRNA [106]. The insulator test vector contains a tissue-specific promoter – Cardiac Actin promoter – that drives expression of a GFP reporter in zebrafish muscle cells upon microinjection. Upstream the muscle-specific promoter the transposon contains the Z48 enhancer, which drives GFP expression in the zebrafish midbrain (**Figure 22**) [15]. Additionally, the vector holds a gateway site where sequences can be easily cloned, between Z48 enhancer and the vector promoter.

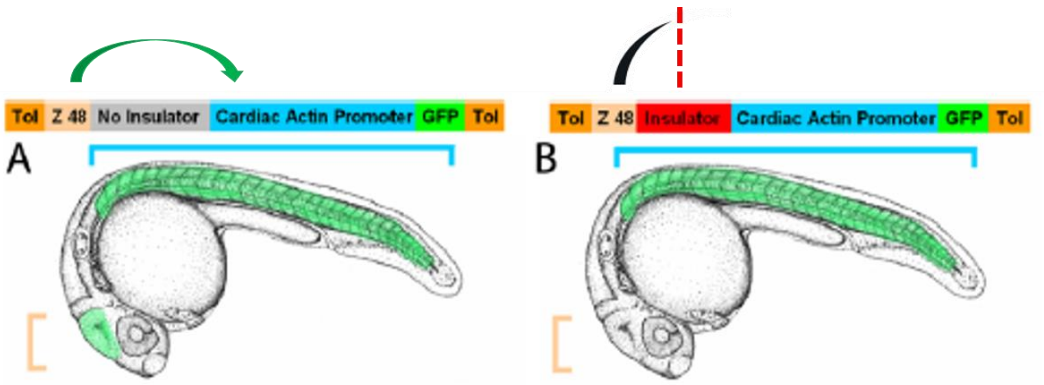

Figure 22. Scheme of the insulator test vector and GFP expression driven in zebrafish. (**A**) Empty insulator test vector. GFP expression is driven in zebrafish somites due tissue-specificity driven by the Cardiac Actin promoter upstream the reporter gene, and in midbrain due to interaction of the Z48 enhancer (midbrain enhancer) with the promoter. (**B**) Insulator test vector containing a cloned insulator upstream the promoter. GFP expression is driven in zebrafish somites, while GFP expression in midbrain is comparatively decreased due to blocking of the interaction between Z48 enhancer and the promoter. Adapted from Bessa, J. et al. (2009).

For these assays, zebrafish embryos were injected with the different insulator test vector containing the respective putative insulators to test, and the empty insulator vector (not containing a cloned insulator) was employed as a negative control in this transgenesis assay. Embryos were documented at 24 hpf (**Figures 23** and **24)** and mean of GFP intensity was quantified by imaging analysis in the midbrain and somites. Because an insulator should impair the expression of GFP detected in the midbrain, while the expression of GFP in somites should be maintained, to access the insulator activity of the tested sequences, the ratio of the mean GFP intensity in the somites divided by the midbrain was calculated, and values were compared to the negative control (**Figure 24**). As positive control, an insulator vector containing the well-known 5'HS4 chicken β-globin insulator was employed in the reporter assay. The 5'HS4

insulator was previously tested in zebrafish [15, 119]. Analysis of the values obtained for the condition used as positive control reveals that all the tested sequences show insulator activity.

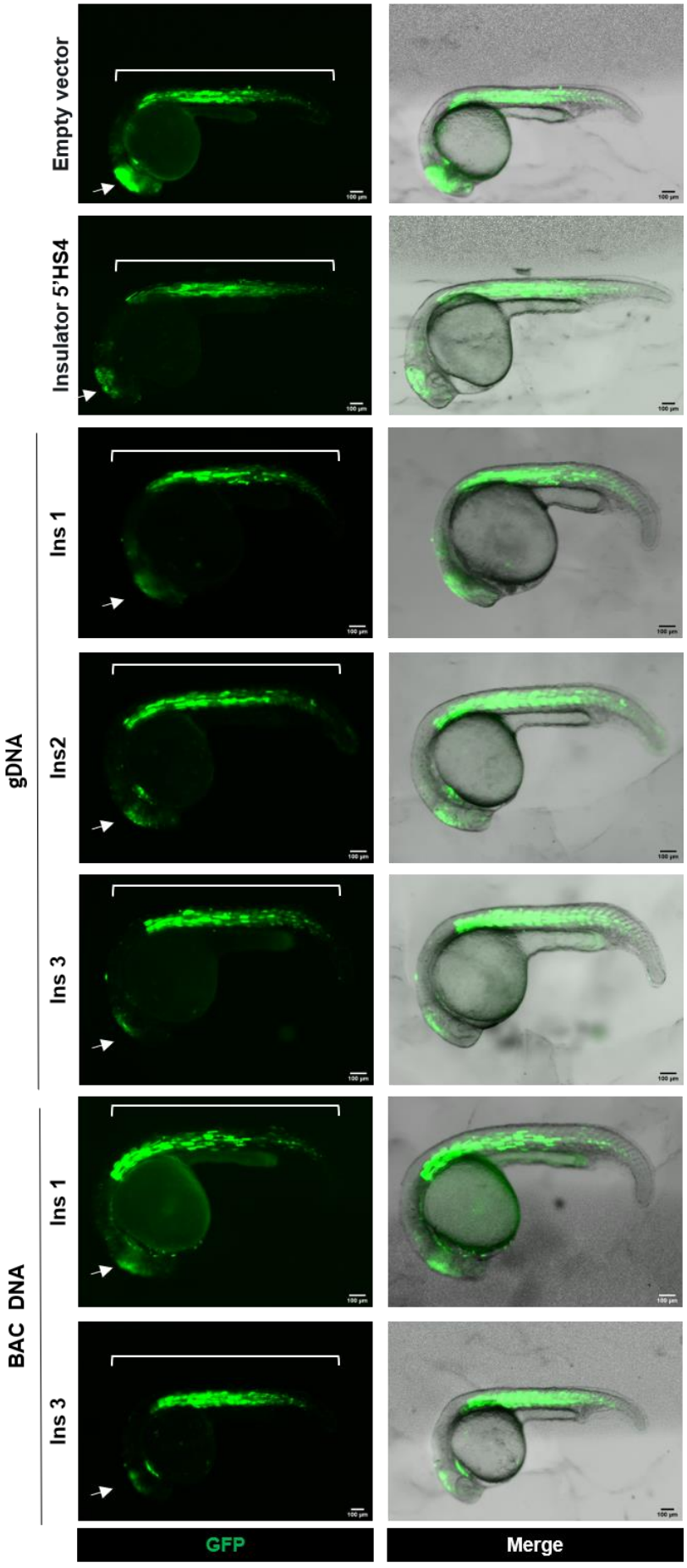

Figure 23. Representative images of 24 hpf zebrafish embryos injected with: empty insulator vector; the insulator vector containing 5'HS4 insulator; and the insulator vector containing ins1 (from BAC and gDNA), ins2 (gDNA) and ins3 (BAC and gDNA). Images acquired in stereomicroscope Leica M205, zoom 5x.

The sequence ins1 showed significant increase of GFP ratio between muscle and midbrain (**Figure 23 and 24**), compared to the negative control, showing its ability to block the interaction of the Z48 enhancer and the Cardiac Actin promoter. Of note, the sequence amplified from human gDNA showed higher insulator activity than the one amplified from BAC DNA. The function of this sequence as an insulator CRE is concordant with the bioinformatic data used in the prediction of PDX1 putative CREs, which shows enrichment of both CTCF and cohesin binding.

The sequence ins2 amplified from human gDNA was also shown to be able to block interaction of the Z48 enhancer with promoter, which led to low GFP expression levels in midbrain of zebrafish. Similarly to ins1, the validation of this sequence as an insulator also corroborates the selection criteria for insulators explained in Chapter 1 (Results). From the Hi-C data, both insulator sequences are the long-range chromatin interaction more distant from the promoter, thus these results confirm our hypothesis that these two sequences might indeed define the border of PDX1 TAD.

Regarding ins3 region, both sequences amplified from human gDNA and BAC DNA were validated as insulators, with a tendency of the first to stronger block the Z48 enhancing action, but not significantly. According to the bioinformatic data used to predict PDX1 putative CREs, the genomic region containing this sequence is enriched in CTCF-binding, although it is not clearly enriched in binding of the cohesin complex, nor is located in a clearly open region of chromatin detected by ATAC-Seq performed in pancreatic islets. Thus, we can hypothesize that the major contribution of ins3 to function as insulator is the CTCF binding.

Altogether, these experimental evidences validated all three selected regions as insulators, using an *in vivo* reporter assay. Next, we aim to investigate the function of these putative CRE in regulating *PDX1* expression at transcriptional level at early developmental stages or in adulthood. An interesting aspect will be to dissect whether these CRE are involved in PDX1 function of pancreas lineage commitment or in mature β-cells.

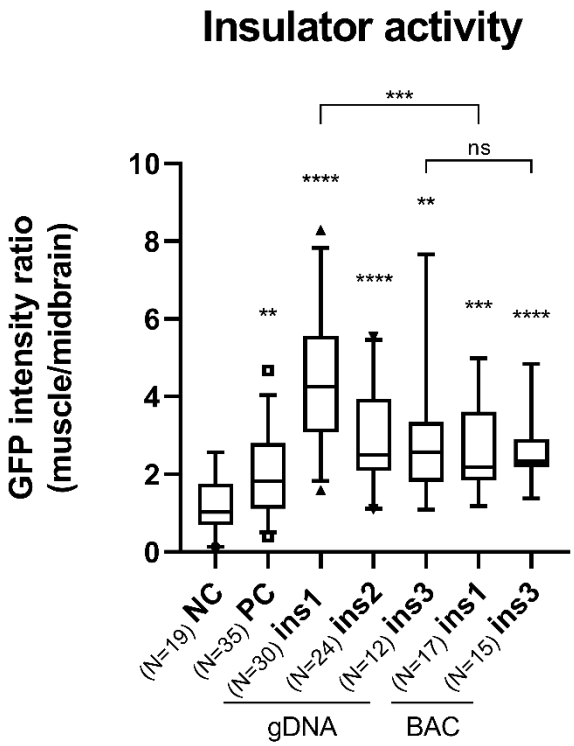

Figure 24. Graphic referring to the ratio between GFP intensity measured in zebrafish somites and in midbrain for each condition tested. “NC” and “PC” refer to negative (Empty vector) and positive control (Insulator 5’HS4), respectively. Statistical analysis was performed with an unpaired T-test with Welch's correction relative to NC. All experimental conditions are statistically significant when comparing to the NC, showing associated P values < 0,05.

Besides validating the sequences in test as insulators, our results show differential levels of insulator activity for sequences amplified from BAC or gDNA DNA. Since we detected variants that differ among the same sequence but amplified from different templates, it is reasonable to hypothesize that the SNVs detected by sequencing are correlated with the activity levels detected. For instance, the statistically significant difference observed between ins1 sequences suggest that SNVs found in the BAC induce a decrease in the ability of that sequence to block enhancer activity. Therefore, it would be interesting to evaluate if the SNVs identified are responsible to modulate the insulator function in the *PDX1* genomic context.

Moreover, the chicken insulator 5’HS4 showed weaker insulator activity in this assay than any of the human sequences tested. The high number of zebrafish individuals injected with the vector containing 5’HS4 analysed in this assay (N=35) provides robustness to our data.

4. Humanization of the zebrafish genome

Besides exploring the potential of individual non-coding sequences contained in *PDX1 locus* as CREs, we have also explored the possibility to study CREs within their original genomic context. Broad cis-regulation analysis was approached through manipulation of human BACs containing the selected regulatory landscape. This last chapter will describe the initial steps of generating a humanized PDX1 zebrafish line, inserting a BACs carrying the human *PDX1* regulatory landscape into the zebrafish genome [56].

1) BAC clone extraction and diagnostic

The BAC containing human *PDX1* regulatory landscape was selected because it spans a long region containing the *PDX1* gene, thus ensuring the whole PDX1 landscape will be present.

First, DNA integrity of PDX1 BAC was validated through electrophoresis analysis. High-weight DNA molecules (as BACs) require particular manipulation to avoid DNA shredding: after extraction the *PDX1* BAC needs to be transferred into another bacteria strain, therefore is crucial that the integrity of the circular molecule is maintained. DNA integrity was confirmed by run the PDX1 BAC plasmid in a 0.8% agarose gel, which reveals the presence of high-weight molecule, in a single band, likely representing the supercoiled structures and others, still located in the pocket of the well. Importantly, no DNA smear as sign of DNA degradation was observed (**Figure 25**).

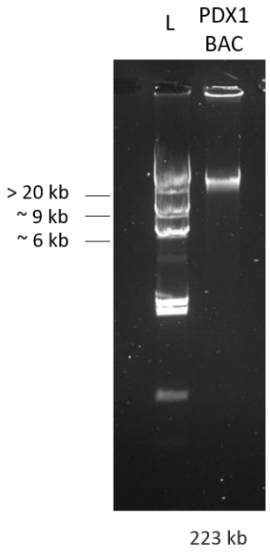

Figure 25. Diagnostic analysis of PDX1 BAC DNA extraction. Electrophoresis of PDX1 BAC plasmid (approximately 200 ng) in 1% (w/v) agarose gel. Symbol “L” represents the Lambda DNA/HindIII Marker and correspondent band sizes.

Further analysis of PDX1 BAC was performed to confirm presence of the correct human genomic sequences selected as putative CREs. The selected sequences were amplified by PCR, followed by run of PCR products on agarose gel. A representative figure of the PCR products is shown in results presented above (see Results; **Figure 7A**). Therefore, PDX1 BAC could be used to reliably represent the human PDX1 *locus*.

2) PDX1 BAC recombineering

To achieve highly efficient transgenesis in zebrafish, PDX1 BAC was manipulated through BAC recombineering. This technique was performed to prepare the BAC DNA so it could be more efficiently integrated in the zebrafish genome upon microinjection – through insertion of *Tol2* transposase recognition sequences [106]. BAC recombineering requires bacterial electroporation of the BAC into a specialized bacterial strain, engineered to express protein for recombination based on homology arms (deriving from the lambda phage), amplification of a linear DNA fragment containing *Tol2* arms, followed by its electroporation, and induction of bacterial recombinase functions.

(1) BAC electroporation

To perform BAC recombineering, we used the recombinogenic bacteria SW102, modified from strain *E. coli* DH10B. Thus, after extraction and purification of plasmid DNA, the PDX1 BAC was electroporated in *E. coli* SW102.

Electroporation of PDX1 BAC into *E. coli* SW102 required several optimizations of the protocol followed [56]. Fresh electrocompetent cells were prepared immediately before electroporation every time, since the efficiency of this type of bacterial transformation is significantly decreased upon cell frost. Aliquots of electrocompetent cells were incubated with purified plasmid DNA (PDX1 BAC) and placed in pre-chilled electroporation cuvettes. Different BAC DNA concentrations were tested, namely 200 ng and 1 ug. The two DNA concentrations were tested in different electroporation conditions, including voltages of 1,2 and 1,8 kV, with constancy of electric capacitance and resistance conditions to 25 mF and 200 Ω, respectively. In every batch of competent bacteria and following BAC electroporation experiments, known concentrations of a supercoiled plasmid were electroporated in parallel, reflecting the conditions under test in each BAC electroporation, namely adjusting volumes. This positive control was used to calculate the competence efficiency of every freshly prepared batch of bacteria.

According to the positive control, higher transformation efficiency of *E. coli* SW102 cells was achieved through use of a voltage of 1,2 kV during electroporation. The transformation efficiency was approximately $9x10^7$ Colony Forming Units (CFU)/ug of DNA, comparing to $1,26x10^7$ CFU/ug when using 1,8 kV. Under the same electroporation conditions, three replicate electroporation experiments with 200 ng and 1 ug of BAC DNA were conducted, resulting in growth of one single colony in one of the three selective agar plates obtained after plating of electroporated cells, for each DNA concentration.

Single colonies grown in selective medium were tested for PDX1 BAC incorporation in two consecutive steps. First, double antibiotic selection was performed to confirm resistance to chloramphenicol (resistance provided by the PDX1 BAC plasmid) and to tetracycline (resistance provided by the strain *E. coli* SW102), illustrated in **Figure 26**. The first electroporation aimed to insert the BAC into the recombinase-competent SW102 strain, maintaining the BAC of interest in single colonies.

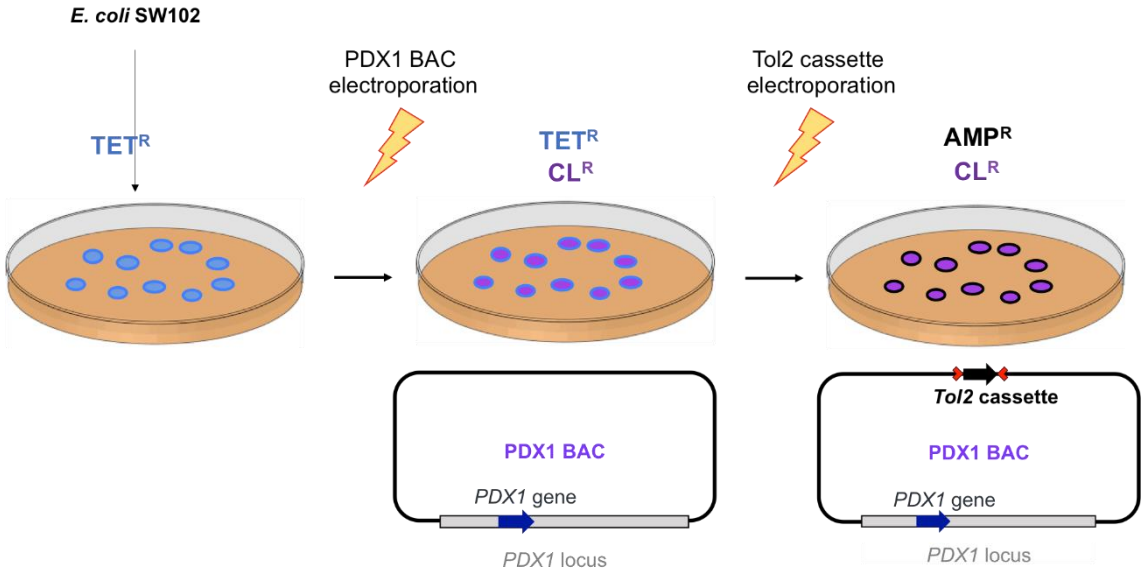

Figure 26. Experimental setup for recombineering of the PDX1 BAC. The bacterial strain E. coli SW102 was plated to select cell clones, represented as single colonies containing resistance to tetracycline (TET^R, in blue). Then, the PDX1 BAC was electroporated into that strain and single colonies of E. coli SW102 containing the BAC were selected by gaining of chloramphenicol resistance (CL^R, in purple). Finally, the Tol2 cassette was electroporated into cells previously selected from single colonies of E. coli SW102:BAC. Colonies containing successful recombined BAC molecules were selected by gaining of ampicillin resistance (AMP^R, in black).

(2) Amplification of Tol2 cassette

To allow efficient BAC transgenesis in zebrafish, the PDX1 BAC was modified through insertion of *Tol2* recognition sequences – the Tol2 cassette.

Amplification of the Tol2 cassette was performed using pCR8GW-iTol2 plasmid as template in a PCR reaction. This vector contains two inverted sequences sufficient for *Tol2* transposase recognition, flanking an ampicillin resistance gene, hereafter referred as Tol2 cassette. Its PCR-based amplification for the recombineering protocol required the design of primers annealing to the edge of the Tol2 cassette with as overhangs 50 bps homology to the targeted site on the backbone of the BAC. The Tol2 cassette was amplified with a proofreading enzyme, followed by electrophoresis of PCR products shown in **Figure 27**. Upon confirmation of PCR amplification of the Tol2 cassette, the template DNA plasmid used for amplification was digested with *DpnI* restriction enzyme. This enzyme belongs to a group of methylation-specific restriction endonucleases, that specifically digests plasmid DNA, because it is methylated. On the contrary, the PCR product is demethylated, therefore will not be digested [120]. Complete digestion of template pCR8GW-iTol2 plasmid was ensured to inhibit arising of false-positive colonies in the BAC recombineering step performed hereafter, deriving from not digested plasmid used as template in the PCR reaction. Preparation of the Tol2 cassette for electroporation in *E. coli* SW102 was concluded with a DNA purification step followed elution in sterile water, because of the downstream applications: in fact, salt containing buffer could impair if not prevent electroporation.

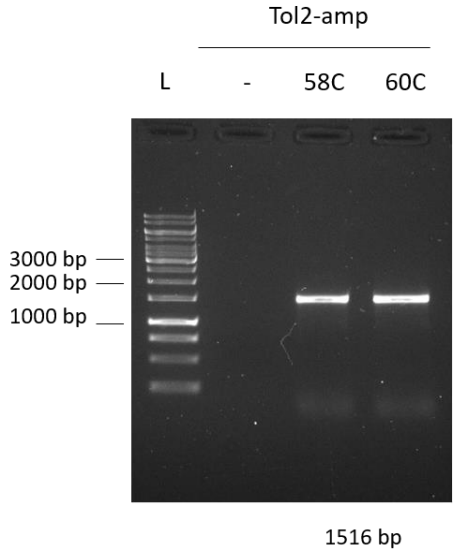

Figure 27. Electrophoresis gel of PCR amplification of Tol2 cassette from pCR8GW-iTol2 plasmid. Symbol "-" refers to PCR negative control (blank), while “L” represents the 1 kb ladder. PCR amplifcation was performed testing two alternative annealing temperatures, 58 and 60 °C. The amplified cassette name is indicated on the top of the gel, while its correspondent size (Tol2 cassette flanked by 50 bp homology arms) is depicted at the bottom.

(3) BAC recombineering of Tol2 cassette

The Tol2 cassette was electroporated in fresh prepared electrocompetent *E. coli* SW102:BAC cells, selected in the previous experimental step (**Figure 27**). During the preparation of electrocompetent cells, recombinase functions of the bacterial strain were inducted by heat-shock, immediately after cell growth until desirable density. Aliquots of electrocompetent cells were incubated with the three different amounts of purified Tol2 cassette – 3,5 ng, 140 ng and 750 ng. Each DNA concentration was electroporated in duplicate. Electroporation shock was conducted in constant conditions, formerly associated to higher transformation efficiency during BAC electroporation – 1,2 kV of voltage, 25 mF of electric capacitance and 200 Ω of resistance.

Insertion of the Tol2 cassette on the backbone of the BAC was achieved for the three concentrations of Tol2 tested in each of the voltages, along cell recovery and bacterial growth in unselective medium for 2 hours. SW102:BAC cells were then plated in medium containing antibiotic. Double antibiotic selection was performed with ampicillin (antibiotic resistance provided by the Tol2 cassette) and chloramphenicol (resistance provided by the PDX1 BAC). In this experimental setting, serial dilution of cells used in the positive control revealed a transformation efficiency of around $8{,}7 \times 10^7$. Cells electroporated with different amounts of the Tol2 cassette were platted separately. A total of 32 single colonies grew on selective medium, of which 19 colonies resulted from the electroporation of 140 ng of Tol2 cassette.

Single colonies grown in selective medium were further tested for successful incorporation of the Tol2 cassette on PDX1 BAC through colony PCR and antibiotic resistance. This selection step, discussed in the following chapter, was particularly crucial due to the possibility of emergence of false-positive colonies.

(4) Selection and preparation of BAC-Tol2 construct

The selection of *E. coli* SW102 cells containing the BAC-Tol2 construct (SW102:BAC-Tol2) was performed by two methods: colony PCR and antibiotic selection.

First, single colonies were tested for presence of the BAC-Tol2 construct by colony PCR. Two PCR amplifications were conducted through use of two pairs of primers – on one hand, we tested the insertion of a DNA sequence on the attB1 site contained in the backbone of the BAC (the target site of recombineering); on the other hand, we tested the insertion of our sequence of interest in the attB1 site, using a specific primer

complementary to the ampicillin resistance gene contained in the Tol2 cassette. The location of the primers designed is depicted in **Figure 28**.

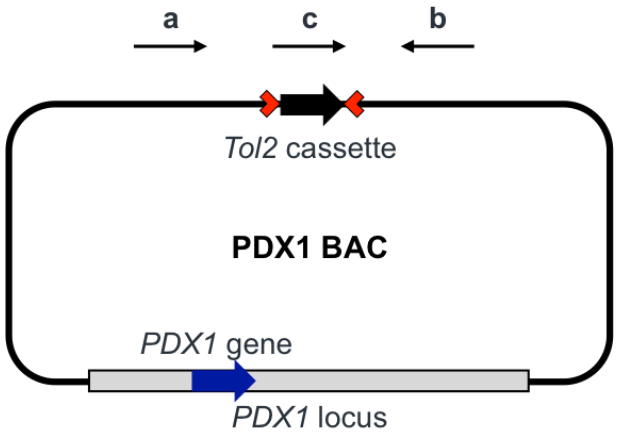

Figure 28. Scheme of the of PDX1 BAC-Tol2 construct and primers used in selection of *E. coli* SW102 cells containing the BAC-Tol2 construct by colony PCR. The two pairs of primers employed were a+b and c+b, which location is represented in the figure.

Results of the colony PCR performed with primers a and b were analysed by electrophoresis in agarose gel (**Figure 29 A**). As these primers were designed flanking the recombineering target site, amplification from a BAC construct without the Tol2 cassette results a small DNA fragment of 148 bp, correspondent to the backbone of the non-recombined plasmid. Amplification from single colonies .1 and .2 resulted in two bands of low molecular weight of high intensity, revealing the presence of non-recombined BAC molecules. The higher molecular weight bands shown in the following gel lanes (colonies .3-.5 and .7) reveal the presence of recombined BAC molecules, which contain the 1416 bp Tol2 cassette in the former attB1 site. Furthermore, the same high molecular weight band is also present in the lanes of colonies .1 and .2, although fainter than the ones obtained from other colonies. These results indicate that the single colonies under test contain different BAC molecules, recombined and non-recombined. Further selection was also performed considering that colonies .3-.5 and .7 present a higher proportion of recombined BAC molecules than colonies .1 and .2, interpreted by the amount of DNA amplified of each size.

Colony PCR results obtained with primers c and b are shown in **Figure 29 B**. This PCR strategy retrieved more specificity in the BAC-Tol2 construct selection, since PCR products were specifically amplified from recombined BAC molecules. Results allowed to confirm that all SW102:BAC-Tol2 colonies tested indeed contain recombined BAC molecules.

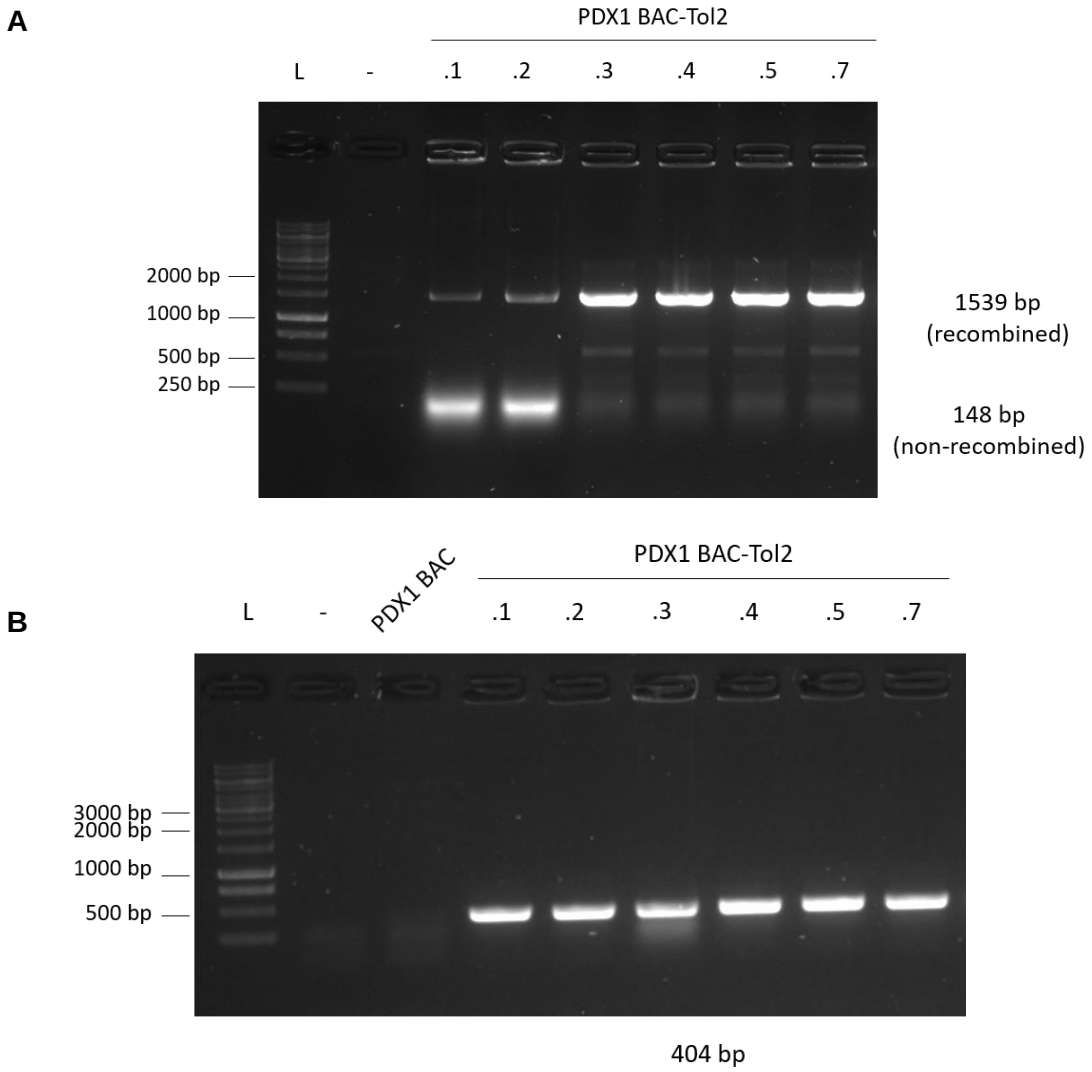

Figure 29. Electrophoresis gel of PCR amplification of the target site of recombineering on the PDX1 BAC from *E. coli* SW102:BAC-Tol2 single colonies. The single colonies are indicated on the top of the gel. Symbol "-" refers to PCR negative control (blank), while "L" represents the 1 kb ladder. (A) Colony PCR products obtained with primers a and b, flanking the recombineering target site. Amplification from a non-recombined BAC molecule results a DNA fragment of 148 bp (backbone of the non-recombined plasmid), while amplification from a recombined BAC molecule results in a DNA fragment of 1539 bp. (B) Colony PCR products obtained with primers c and b, designed to amplify part of the Tol2 cassette inserted in a BAC molecule. Amplification of a single band corresponding to the junction between the Tol2 cassette and a downstream sequence of the recombineering target site allows to detected successful recombineering. Size of the amplified sequence is depicted at the bottom.

Moreover, selection of *E. coli* SW102:BAC-Tol2 colonies was complemented by an examination of antibiotic resistance. This strategy was relevant not only to confirm presence recombined BAC molecules inside *E. coli* SW102 bacteria, but also to detect the growth of false-positive colonies upon electroporation of the Tol2 cassette. In the end

of the Tol2 recombineering protocol, three different DNA combinations can be contained in the cells grown on selective plates containing ampicillin and chloramphenicol. One is the desired recombined BAC molecule, comprising the resistance genes for both antibiotics in its backbone due to successful recombineering of the Tol2 cassette; other is the non-recombined BAC molecule, only conferring resistance to chloramphenicol; finally, the pCR8GW-iTol2 plasmid used as template for amplification of the Tol2 cassette could also be contained within these cells, conferring resistance to ampicillin. As mentioned above, complete digestion of pCR8GW-iTol2 plasmid from the PCR reaction by *DpnI* restriction enzyme was required in order to prevent growth of false-positive colonies. Even though the restriction reaction was set up accordingly to the recommended protocol, we could not ensure a 100 % effective restriction *a priori*. Thus, electroporation of a DNA sample containing the Tol2 cassette along with a small portion of pCR8GW-iTol2 plasmid would be sufficient to allow the growth of *E. coli* SW102 cells on LB-chloramphenicol-ampicillin plates. These false-positive colonies containing both non-recombined BAC molecules and pCR8GW-iTol2 plasmid should then be discarded. To do so, *E. coli* SW102:BAC-Tol2 colonies were streaked in LB-spectinomycin plates – additional antibiotic resistance cassette specific to this plasmid. Streak of the 32 single colonies grown upon electroporation in selective plates containing spectinomycin allowed to identify one false-positive. From the 31 colonies left, maintenance of the PDX1 BAC on the correct bacterial strain was ensured by streaking in double selective agar plates containing, chloramphenicol and ampicillin. Single colonies grown on those plates were then used to extract the PDX1 BAC-Tol2 construct.

3) Humanizing the zebrafish genome: PDX1 BAC transgenesis

(5) pdx1 zebrafish mutant line

As part of this thesis, a zebrafish mutant line for the *pdx1 locus* was reared in the animal facility - pdx1$^{sa280/sa280}$. Professor Robin A. Kimmel and colleagues, who described this zebrafish mutant line generated by the Zebrafish Mutation Project, provided us with offspring of an incross between two heterozygous animals for the *pdx1* mutation [93].

The zebrafish pdx1$^{sa280/sa280}$ line is characterised by a mutant allele containing a premature stop at codon (Y37X), located within the N-terminal protein transactivation domain. Differently from humans, zebrafish embryos containing the null mutation sa280 on the *pdx1 locus* in the homozygous state reach adulthood, developing a pancreas. However, the mutants show decreased viability and a specific pancreatic phenotype,

characterised by reduced number of β-cells, low levels of insulin and perturbed differentiation of acinar cells. Furthermore, zebrafish *pdx1* mutants respond to drug treatment used in human diabetic patients, which highlights its potential use as a genetic model of diabetes [93].

The establishment of the zebrafish *pdx1* mutant line was conducted in order to test the rescue of loss-of-function upon transgenesis with the PDX1 BAC. We aim to determine whether the human *PDX1* gene and its corresponding regulatory landscape can rescue the loss-of-function of its zebrafish ortholog. Thus, the human PDX1 BAC will be used to perform transgenesis in a *pdx1* null background, provided by this zebrafish mutant line. If the rescue is feasible, this will be an excellent model to study the phenotypic consequences of non-coding regulatory mutations in the human genomic landscape of PDX1.

The offspring of heterozygous *pdx1* zebrafish mutants was raised until sexual maturity, so they could be genotyped for the mutation. Each animal was genotyped through outcross with a zebrafish WT line and DNA extraction from embryos. Upon PCR amplification of the zebrafish *pdx1 locus*, a digestion reaction with *DraI* restriction enzyme was performed (**Figure 30 A**). The rationale of this genotyping strategy lies in the fact that the null mutation contained in the *pdx1 locus* of the mutants creates a restriction *DraI* site, which is absent in non-mutated *pdx1* alleles. Thus, after *DraI* digestion and electrophoresis, DNA extracted from embryos containing the mutant allele is detected by the emergence of a lower molecular weight DNA band, correspondent to a DNA fragment of 334 bp (right side of **Figure 30 A**). Lanes 1 to 4 refers to DNA extracted from batches of embryos holding the *pdx1* mutant allele, which allows to identify progenitors as *pdx1* mutants. **Figure 30 B** shows the difference at the sequence level among WT and *pdx1* mutants, and corresponding *DraI* restriction site.

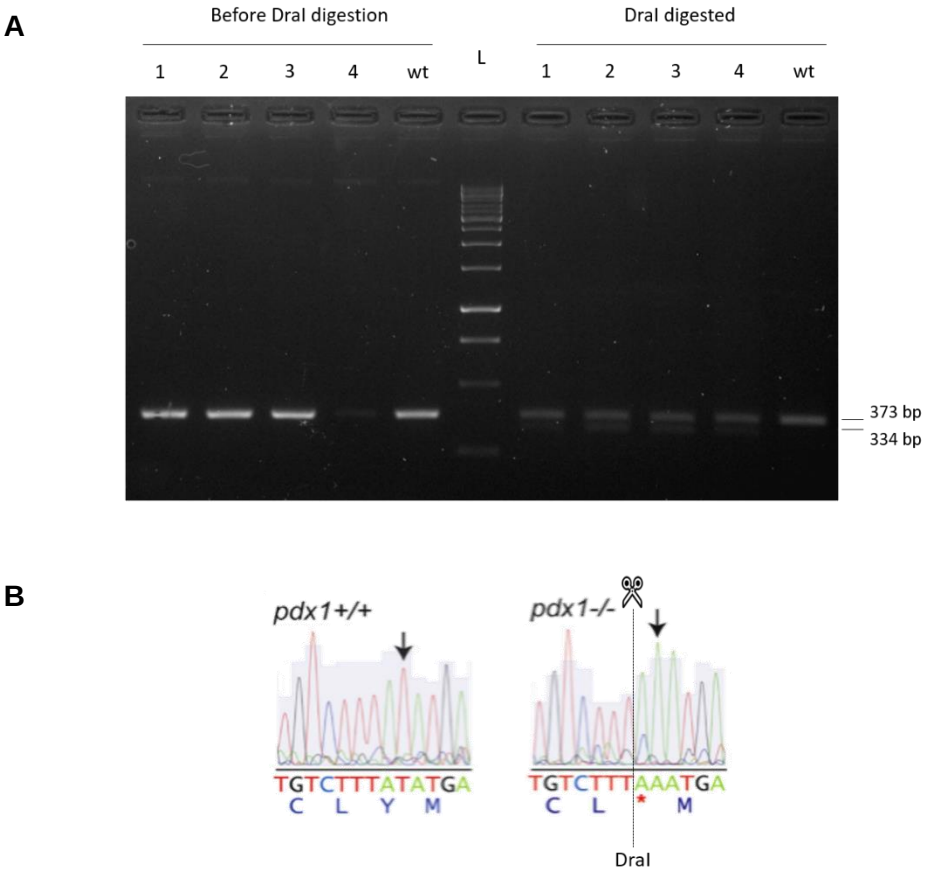

Figure 30. Genotyping of zebrafish pdx1 mutants. (A) Electrophoresis gel of PCR amplification of pdx1 *locus* from DNA extracted from batches of 8 zebrafish embryos (left side, "Before *DraI* digestion") and the same PCR product upon digestion with *DraI* restriction enzyme (right side, "*DraI* digested"). Embryos correspond to the offspring of an outcross between putative carriers of the pdx1 mutant allele with wt animals. Sizes for wt and pdx1 mutant allele upon digestion with *DraI* are presented on the right side of the gel. Symbol "wt" refers to genotyping negative control performed with DNA from wt embryos, while "L" represents the 1 kb ladder. (B) Representation of sequencing of genomic DNA obtained from zebrafish wt animals (pdx1+/+) and pdx1 homozygous mutants (pdx1-/-). Zebrafish pdx1 mutants hold a restriction site for *DraI* enzyme in that *locus*.

(6) BAC microinjection and toxicity evaluation

The PDX1 BAC-Tol2 construct was used to injected one-cell stage zebrafish embryos. In this first approach, microinjection was performed in wt embryos to determine the range of BAC DNA amount in which we detect efficiency of transgenesis without compromising animal vitality. Since microinjection in zebrafish embryos is often performed with vectors of much smaller size than a BAC, we aimed to assess whether the amount of DNA injected should reflect the size of the transposon that will be integrated in the zebrafish genome.

Three concentrations of PDX1 BAC were injected, namely 50, 150 and 500 ng/uL of DNA, each one tested with and without co-microinjection of *Tol2* mRNA at 25 ng/uL [106]. Upon microinjection, the rate of mortality was examined at 10 dpf (**Table 8)**. The highest mortality rate was found to be correlated with the highest amount of BAC DNA microinjected. Whereas 500 ng appears indeed to be too toxic, a clear difference between 50 and 150 ng was not observed, suggesting this as optimal range.

Table 8. Experimental conditions of PDX1 BAC microinjection in one-cell zebrafish embryos and mortality rates. Concentrations of BAC DNA as well as *Tol2* mRNA are indicated.

PDX1 BAC DNA (ng/uL)	*Tol2* mRNA (ng/uL)	Mortality at 10 dpf (%)
50	-	30
	25	30
150	-	20
	25	10-50
500	-	80
	25	90

(7) Genotyping BAC transgenics

In order to assess transgenesis efficiency values, we employed a genotyping strategy to identify BAC transgenic animals. In this experiment, we used batches of four zebrafish animals microinjected with the PDX1 BAC to extract DNA, grouped according to each condition tested in microinjection. The larvae were collected at 10 dpf to avoid contamination from non-integrated DNA fragments. We used primers designed specifically to the promoter of the human *PDX1 locus* (PDX1 P), thus determining whether integration of the human transposon into the zebrafish genome occurred.

We first tested the specificity of our primers within mixes of human DNA, both from PDX1 BAC and gDNA, and zebrafish DNA (ZF) (**Figure 31**). Using BAC DNA as template for PCR amplification, a higher yield of PCR product was obtained comparing to gDNA. This difference seems to reflect the presence of a higher number of copies of the template in the BAC DNA, when using the same total amount of BAC DNA and gDNA. Analysis of PCR products allows to conclude that mixing of zebrafish DNA with human DNA sustains primer specificity, since PCR amplification from that DNA mix does not show additional PCR products.

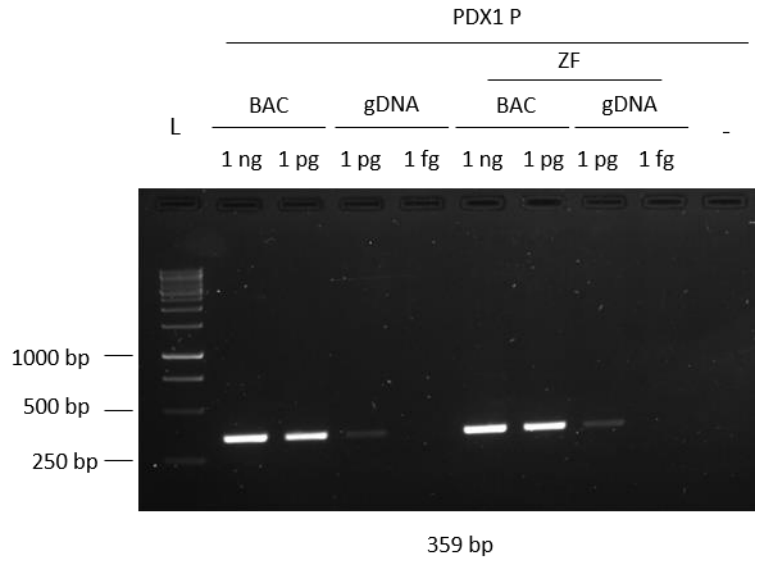

Figure 31. Electrophoresis gel of PCR amplification of human PDX1 promoter region (PDX1 P). DNA templates used are depicted above the gel, which were the following: PDX1 BAC DNA (BAC), human gDNA (gDNA), zebrafish DNA (ZF) mixed with BAC and ZF DNA mixed with human gDNA. ZF DNA corresponds to approximately 10 ng. Each template of human DNA was tested in two concentrations: 1 ng and 1 pg for BAC DNA and 1 pg and 1 fg for gDNA. Symbol "-" refers to PCR negative control (blank); while "L" represents the 1 kb ladder. Size of the amplified region is depicted at the bottom.

After confirming primers specificity, we investigated which amount of human DNA that could be used in PCR in order to detect a successful integration of the BAC in the zebrafish genome. To evaluate PCR sensibility, we estimated the amount of BAC DNA or human gDNA employed in PCR that should mimic the event of a single molecule insertion of the human PDX1 BAC into the zebrafish genome. Taking into consideration the full size of the zebrafish genome and the size of the PDX1 BAC, we estimated that the ratio between total amount of BAC DNA and ZF DNA is around $1:10^4$ (see Materials and Methods, chapter 9. c) (3)). In other words, when performing a PCR reaction with 10 ng (10000 pg) of ZF DNA, 1 pg of PDX1 BAC in that reaction should represent 1 copy of the BAC integrated in the zebrafish genome. We performed the previous PCR reaction using as template the DNA extracted from batches of four zebrafish animals at 10 dpf microinjected with the PDX1 BAC. As PCR controls, we used human gDNA, ZF DNA and ZF DNA mixed with either human gDNA or the BAC (mimicking 1 copy of BAC per 1, 100 or 1000 copies of the zebrafish genome, respectively). A first PCR reaction was performed with thirty-five amplification cycles, which results were able to detect the human PDX1 BAC only in "1 BAC + ZF" (**Figure 32 A**; see Materials and Methods, chapter 9. c) (3)). This indicates that the resolution of the genotyping PCR was sufficient to detect one molecule of PDX1 BAC per zebrafish genome.

Thus, to enhance the sensibility of the PCR even further (**Figure 32 B**, lane "1 BAC + 100 ZF"), the PCR was performed by increasing the number of amplification cycles to

forty. By this optimization, in both controls, “1 BAC + 1 ZF” and “1 BAC + 100 ZF”, the amplification of the human PDX1 BAC was detected, indicating that the resolution achieved ranges between one and one hundredth BAC copy per genome. With these conditions, the desired amplicon was detected in two distinct batches of injected embryos - the two correct bands are identified by the yellow arrows (**Figure 32 B,** lanes “BAC 50 ng .4” and “BAC 150 ng .3”). Concomitantly, various unspecific PCR amplicons arose, which is expected with such high number of cycles. The two positive batches for the PCR derived from embryos microinjected with 50 ng of BAC and 150 ng of BAC, confirming that both experimental conditions might be successful. Further optimization of the PCR protocol, as well as confirmation by Sanger sequencing of the expected size bands will be performed in the near future. As preliminary data, our results suggest that microinjection of 50 ng of BAC is sufficient to allow transgenesis, since we were able to amplify the correct sequence from the DNA from batches of embryos injected with this amount of BAC (**Figure 32 B**, lane “BAC 50 ng .4”).

Determination of significant values of transgenesis efficiency is crucial to determine the use of the BAC injected animals as founders of a stable transgenic line. To establish the PDX1 transgenic line, we planned to inject with the highest amount of BAC DNA (150ng) in order to maximize the transgenesis efficiency, since the toxicity range is similar.

The generation of the PDX1 transgenic line will be used to dissect *PDX1* transcriptional regulation and represent a humanized model, which is easy to edit genetically, thus allowing deleting CRE to investigate their biological impact on *PDX1* expression.

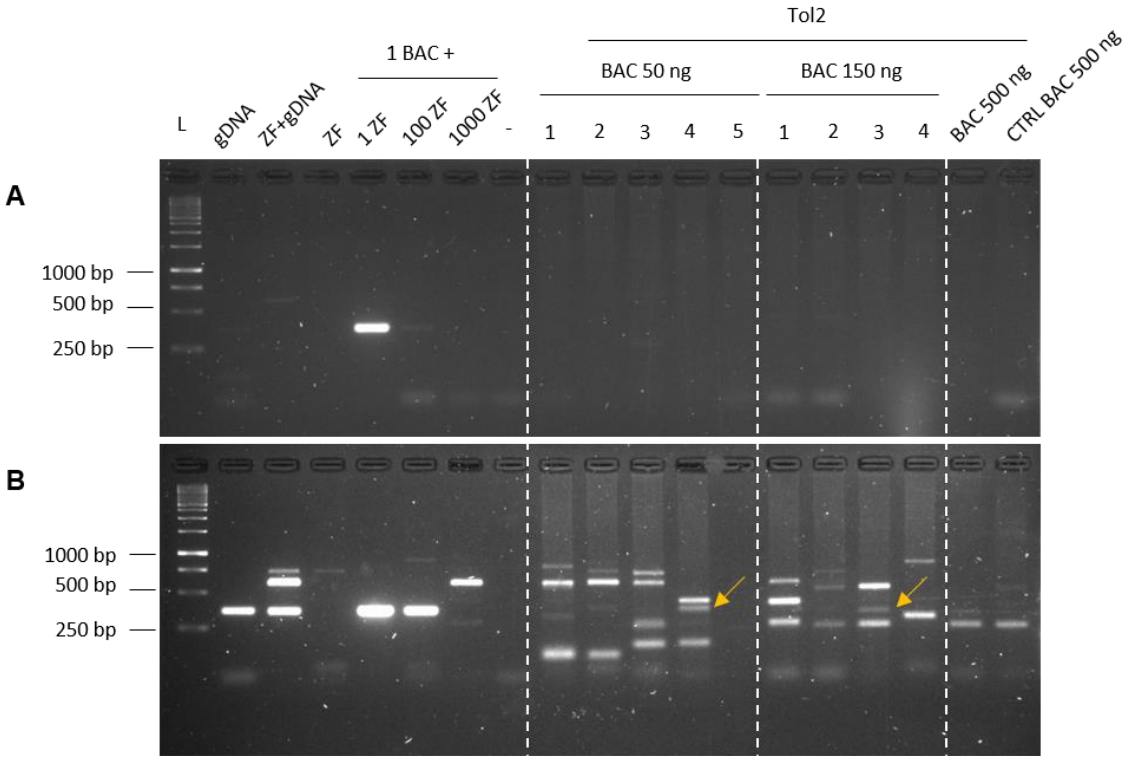

Figure 32. Electrophoresis gels of PCR of the human PDX1 *locus* from zebrafish animals microinjected with the PDX1 BAC. PCR amplification was performed using batches of four zebrafish individuals at 10dpf (identified with numbers between 1 and 5) from the three BAC concentrations, along with *Tol2* mRNA tested in microinjection, or with BAC DNA only (without *Tol2*), named "CTRL BAC 500ng". PCR controls (left side of both gels) were performed through amplification of: human gDNA alone (gDNA); zebrafish DNA alone (ZF); a mixture of ZF with human gDNA (ZF+gDNA); and mixtures of ZF with the BAC in different concentrations, identified at the top of the gel (corresponding to 1 BAC copy in 1, 100 and 1000 molecules of the zebrafish genome, respectively). Symbol "-" refers to PCR negative control (blank), while "L" represents the 1 kb ladder. (**A**) PCR results obtained with 35 cycles of amplification. (**B**) PCR results obtained with 40 cycles of amplification.

General conclusions and future perspectives

Complex diseases, such as T2D, have been increasing their incidence in worldwide human populations. Particularly, more than 422 million adults have been reported as diabetic patients in 2016 [68]. As a multifactorial disease, the pathogenesis of diabetes outcomes from risk factors including genetic predisposition, aging and environmental factors [65-67]. Moreover, similarly to cancer, diabetes diagnosis can be accompanied by failure in different organ systems, for instance the central nervous system [121]. GWAS have been successfully employed to characterise genetic variants associated to disease, which detected several human polymorphisms associated to T2D predisposition [122]. T2D-associated polymorphisms are frequently located within non-coding regions characterised by epigenetic marks of CREs, namely enhancers; furthermore, analysis of eQTLs allowed to assign an impact of several of those polymorphisms in gene expression dysregulation [7, 123]. Co-localization of T2D-associated polymorphisms and gene regulatory elements, such as enhancers, suggests an impact on transcriptional regulation [34-36].

The main focus of this thesis was to explore the regulatory landscape of *PDX1*, a master regulator of pancreatic development and function. Starting from the choice of *PDX1* as our *locus* of interest, we analysed the genomic context of a T2D-associated SNP identified by GWAS, which is contained within the gene *locus*. Based on reports showing the localization of T2D-associated SNP in genomic regions enriched in epigenetic marks used to define enhancers, we defined a putative enhancer sequence containing the mentioned polymorphism. Together with other two putative enhancers that were defined using chromatin accessibility and conformation data from human pancreatic islets, we tested the ability of those sequences to enhance the expression of reporter genes in the endocrine pancreas and pancreatic progenitors through zebrafish transgenesis. Moreover, we also defined putative insulator sequences flanking *PDX1* regulatory landscape, based on chromatin interactions detected by Hi-C data in human pancreatic islets. On one hand, our results showed that the enhancer eSNP, the sequence overlapping with the T2D-associated SNP, is able to drive GFP expression in the pancreatic domain of zebrafish embryos, although very mildly, while en1 and en2 sequences were not, therefore they are unable to function as enhancers, at least in this reporter assay. On the other hand, our results showed that all the three putative CREs successfully function as insulators in an *in vivo* enhancer blocking assay in zebrafish.

Overall, with these results, we were able to map and validate new CREs in the regulatory landscape of *PDX1*.

Besides screening for CREs contained within *PDX1 locus* through zebrafish transgenesis assays, the novelty of the project presented in this thesis relies on the study the full regulatory landscape of this master regulator of the pancreas. While the transgenesis assays to validate CREs were performed by isolating the genomic sequence and exploring its impact on the expression of a reporter gene, we will further study *PDX1* CREs without disturbing their original genomic context. For this purpose, a *PDX1* BAC was selected and evaluated as a reliable representation of the human *PDX1 locus*. To introduce the full human *PDX1 locus* in the zebrafish genome, we took advantage of the established protocol of BAC recombineering, by which we inserted a Tol2 cassette in the *PDX1* BAC, thus improving transgenesis efficiency upon microinjection in zebrafish. While the *PDX1* BAC was engineered for efficient transgenesis, a zebrafish mutant line for the *pdx1 locus* was reared in the animal facility. Zebrafish *pdx1* mutants provide an *in vivo* tool that will further be used to test the ability of *PDX1* BAC to rescue loss-of-function of the zebrafish ortholog upon transgenesis. Preliminary data of *PDX1*-BAC-Tol2 injection in *pdx1* heterozygotes, as well as in wt zebrafish, showed a transgenesis efficiency between 5 and 20%. Injected fish are growing, and when reached the sexual maturity, they will be screened for germline transmission to identify founders for establishing the *PDX1*-zebrafish line.

As future perspectives, we aim to develop a transgenic line of animals containing *PDX1 locus*. While maintaining the original structure of its regulatory landscape, we can study human cis-regulation of *PDX1 in vivo*. We aim to assess the impact of the SNP mapped in *PDX1* regulatory elements previously associated to T2D, as a potential cause for the disruption of normal pancreatic function. This will be done using CRISPR-Cas9 site directed mutagenesis system. In addition, we want to identify novel polymorphisms that could potentially impair proper development of the pancreas. Using zebrafish as model organism, PDX1 BAC transgenics will constitute an *in vivo* “tool” that could be used to approach both single and congregated impact of non-coding polymorphisms on diabetes predisposition.

References

1. Raciti, G.A., M. Longo, L. Parrillo, M. Ciccarelli, P. Mirra, P. Ungaro, P. Formisano, C. Miele, and F. Beguinot, *Understanding type 2 diabetes: from genetics to epigenetics.* Acta Diabetol, 2015. **52**(5): p. 821-7.
2. Lodish, U.H., *Molecular Cell Biology*. 2016: W.H. Freeman.
3. Huret, J.L., M. Ahmad, M. Arsaban, A. Bernheim, J. Cigna, F. Desangles, J.C. Guignard, M.C. Jacquemot-Perbal, M. Labarussias, V. Leberre, A. Malo, C. Morel-Pair, H. Mossafa, J.C. Potier, G. Texier, F. Viguie, S. Yau Chun Wan-Senon, A. Zasadzinski, and P. Dessen, *Atlas of genetics and cytogenetics in oncology and haematology in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D920-4.
4. Klemm, S.L., Z. Shipony, and W.J. Greenleaf, *Chromatin accessibility and the regulatory epigenome.* Nature Reviews Genetics, 2019. **20**(4): p. 207-220.
5. Michalak, E.M., M.L. Burr, A.J. Bannister, and M.A. Dawson, *The roles of DNA, RNA and histone methylation in ageing and cancer.* Nat Rev Mol Cell Biol, 2019.
6. Levine, M., C. Cattoglio, and R. Tjian, *Looping back to leap forward: transcription enters a new era.* Cell, 2014. **157**(1): p. 13-25.
7. Pasquali, L., et al., *Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants.* Nat Genet, 2014. **46**(2): p. 136-143.
8. Trynka, G., *Enhancers looping to target genes.* Nat Genet, 2017. **49**(11): p. 1564-1565.
9. Maston, G.A., S.K. Evans, and M.R. Green, *Transcriptional regulatory elements in the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 29-59.
10. Bessa, J., M. Luengo, S. Rivero-Gil, A. Ariza-Cosano, A.H. Maia, F.J. Ruiz-Ruano, P. Caballero, S. Naranjo, J.J. Carvajal, and J.L. Gomez-Skarmeta, *A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates.* Genome Res, 2014. **24**(3): p. 487-95.
11. Furlong, E.E.M. and M. Levine, *Developmental enhancers and chromosome topology.* Science, 2018. **361**(6409): p. 1341-1345.
12. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a β-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2): p. 299-308.
13. Lettice, L.A., S.J. Heaney, L.A. Purdie, L. Li, P. de Beer, B.A. Oostra, D. Goode, G. Elgar, R.E. Hill, and E. de Graaff, *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Hum Mol Genet, 2003. **12**(14): p. 1725-35.
14. Malik, S. and R.G. Roeder, *The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation.* Nat Rev Genet, 2010. **11**(11): p. 761-72.
15. Bessa, J., J.J. Tena, E. de la Calle-Mustienes, A. Fernandez-Minan, S. Naranjo, A. Fernandez, L. Montoliu, A. Akalin, B. Lenhard, F. Casares, and J.L. Gomez-Skarmeta, *Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish.* Dev Dyn, 2009. **238**(9): p. 2409-17.
16. West, A.G., M. Gaszner, and G. Felsenfeld, *Insulators: many functions, many mechanisms.* Genes Dev, 2002. **16**(3): p. 271-88.
17. Ciavatta, D., S. Kalantry, T. Magnuson, and O. Smithies, *A DNA insulator prevents repression of a targeted X-linked transgene but not its random or imprinted X inactivation.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(26): p. 9958-9963.
18. Miguel-Escalada, I., et al., *Human pancreatic islet 3D chromatin architecture provides insights into the genetics of type 2 diabetes.* bioRxiv, 2018: p. 400291.

19. Hansen, A.S., C. Cattoglio, X. Darzacq, and R. Tjian, *Recent evidence that TADs and chromatin loops are dynamic structures.* Nucleus, 2018. **9**(1): p. 20-32.
20. Hansen, A.S., I. Pustova, C. Cattoglio, R. Tjian, and X. Darzacq, *CTCF and cohesin regulate chromatin loop stability with distinct dynamics.* eLife, 2017. **6**: p. e25776.
21. Hnisz, D., D.S. Day, and R.A. Young, *Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control.* Cell, 2016. **167**(5): p. 1188-1200.
22. Harmston, N., E. Ing-Simmons, G. Tan, M. Perry, M. Merkenschlager, and B. Lenhard, *Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation.* Nat Commun, 2017. **8**(1): p. 441.
23. Mularoni, L., M. Ramos-Rodriguez, and L. Pasquali, *The Pancreatic Islet Regulome Browser.* Front Genet, 2017. **8**: p. 13.
24. Chung, H.-R., I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A.E. Ehrenhofer-Murray, S.R. Sperling, and M. Vingron, *The effect of micrococcal nuclease digestion on nucleosome positioning data.* PloS one, 2010. **5**(12): p. e15754-e15754.
25. Sullivan, A.M., K.L. Bubb, R. Sandstrom, J.A. Stamatoyannopoulos, and C. Queitsch, *DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants.* Current Plant Biology, 2015. **3-4**: p. 40-47.
26. Gaulton, K.J., T. Nammo, L. Pasquali, J.M. Simon, P.G. Giresi, M.P. Fogarty, T.M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K.L. Mohlke, J.D. Lieb, and J. Ferrer, *A map of open chromatin in human pancreatic islets.* Nat Genet, 2010. **42**(3): p. 255-9.
27. Buenrostro, J.D., B. Wu, H.Y. Chang, and W.J. Greenleaf, *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Current protocols in molecular biology, 2015. **109**: p. 21.29.1-21.29.9.
28. Parker, S.C., M.L. Stitzel, D.L. Taylor, J.M. Orozco, M.R. Erdos, J.A. Akiyama, K.L. van Bueren, P.S. Chines, N. Narisu, N.C.S. Program, B.L. Black, A. Visel, L.A. Pennacchio, F.S. Collins, A. National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, and N.C.S.P. Authors, *Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17921-6.
29. Sati, S. and G. Cavalli, *Chromosome conformation capture technologies and their impact in understanding genome function.* Chromosoma, 2017. **126**(1): p. 33-44.
30. Dekker, J., K. Rippe, M. Dekker, and N. Kleckner, *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
31. Dekker, J., *The three 'C' s of chromosome conformation capture: controls, controls, controls.* Nat Methods, 2006. **3**(1): p. 17-21.
32. Matelot, M. and D. Noordermeer, *Determination of High-Resolution 3D Chromatin Organization Using Circular Chromosome Conformation Capture (4C-seq).* Methods Mol Biol, 2016. **1480**: p. 223-41.
33. Lieberman-Aiden, E., N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, and J. Dekker, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.
34. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nat Rev Genet, 2015. **16**(4): p. 197-212.
35. Flannick, J. and J.C. Florez, *Type 2 diabetes: genetic data sharing to advance complex disease research.* Nat Rev Genet, 2016. **17**(9): p. 535-49.

36. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science (New York, N.Y.), 2012. **337**(6099): p. 1190-1195.
37. Mahajan, A., et al., *Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes.* Nat Genet, 2018. **50**(4): p. 559-571.
38. Wang, X., L. He, S.M. Goggin, A. Saadat, L. Wang, N. Sinnott-Armstrong, M. Claussnitzer, and M. Kellis, *High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human.* Nat Commun, 2018. **9**(1): p. 5380.
39. Cebola, I., S.A. Rodriguez-Segui, C.H. Cho, J. Bessa, M. Rovira, M. Luengo, M. Chhatriwala, A. Berry, J. Ponsa-Cobas, M.A. Maestro, R.E. Jennings, L. Pasquali, I. Moran, N. Castro, N.A. Hanley, J.L. Gomez-Skarmeta, L. Vallier, and J. Ferrer, *TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors.* Nat Cell Biol, 2015. **17**(5): p. 615-626.
40. Matsuda, H., *Zebrafish as a model for studying functional pancreatic beta cells development and regeneration.* Dev Growth Differ, 2018. **60**(6): p. 393-399.
41. Andersen, D.K., M. Korc, G.M. Petersen, G. Eibl, D. Li, M.R. Rickels, S.T. Chari, and J.L. Abbruzzese, *Diabetes, Pancreatogenic Diabetes, and Pancreatic Cancer.* Diabetes, 2017. **66**(5): p. 1103-1110.
42. Burlison, J.S., Q. Long, Y. Fujitani, C.V. Wright, and M.A. Magnuson, *Pdx-1 and Ptf1a concurrently determine fate specification of pancreatic multipotent progenitor cells.* Dev Biol, 2008. **316**(1): p. 74-86.
43. Bakhti, M., A. Böttcher, and H. Lickert, *Modelling the endocrine pancreas in health and disease.* Nature Reviews Endocrinology, 2018.
44. Jennings, R.E., A.A. Berry, J.P. Strutt, D.T. Gerrard, and N.A. Hanley, *Human pancreas development.* Development, 2015. **142**(18): p. 3126-37.
45. Bastidas-Ponce, A., K. Scheibner, H. Lickert, and M. Bakhti, *Cellular and molecular mechanisms coordinating pancreas development.* Development, 2017. **144**(16): p. 2873-2888.
46. Marty-Santos, L. and O. Cleaver, *Pdx1 regulates pancreas tubulogenesis and E-cadherin expression.* Development (Cambridge, England), 2016. **143**(1): p. 101-112.
47. McCracken, K.W. and J.M. Wells, *Molecular pathways controlling pancreas induction.* Semin Cell Dev Biol, 2012. **23**(6): p. 656-62.
48. Larsen, H.L. and A. Grapin-Botton, *The molecular and morphogenetic basis of pancreas organogenesis.* Semin Cell Dev Biol, 2017. **66**: p. 51-68.
49. Shih, H.P., J.L. Kopp, M. Sandhu, C.L. Dubois, P.A. Seymour, A. Grapin-Botton, and M. Sander, *A Notch-dependent molecular circuitry initiates pancreatic endocrine and ductal cell differentiation.* Development (Cambridge, England), 2012. **139**(14): p. 2488-2499.
50. Qu, X., S. Afelik, J.N. Jensen, M.A. Bukys, S. Kobberup, M. Schmerr, F. Xiao, P. Nyeng, M. Veronica Albertoni, A. Grapin-Botton, and J. Jensen, *Notch-mediated post-translational control of Ngn3 protein stability regulates pancreatic patterning and cell fate commitment.* Dev Biol, 2013. **376**(1): p. 1-12.
51. Eames Nalle, S.C., K.F. Franse, and M.D. Kinkel, *Chapter 11 - Analysis of pancreatic disease in zebrafish*, in *Methods in Cell Biology*, H.W. Detrich, M. Westerfield, and L.I. Zon, Editors. 2017, Academic Press. p. 271-295.
52. Kimmel, R.A. and D. Meyer, *Zebrafish pancreas as a model for development and disease.* Methods Cell Biol, 2016. **134**: p. 431-61.
53. Westerfield, M., *The zebrafish book : a guide for the laboratory use of zebrafish (Danio rerio)*. 2007.

54. Giraldo, P. and L. Montoliu, *Size matters: use of YACs, BACs and PACs in transgenic animals.* Transgenic research, 2001. **10**: p. 83-103.
55. Chen, X., D. Gays, and M.M. Santoro, *Transgenic Zebrafish.* Methods Mol Biol, 2016. **1464**: p. 107-114.
56. Suster, M.L., G. Abe, A. Schouw, and K. Kawakami, *Transposon-mediated BAC transgenesis in zebrafish.* Nat Protoc, 2011. **6**(12): p. 1998-2021.
57. O'Hare, E.A., L.M. Yerges-Armstrong, J.A. Perry, A.R. Shuldiner, and N.A. Zaghloul, *Assignment of Functional Relevance to Genes at Type 2 Diabetes-Associated Loci Through Investigation of beta-Cell Mass Deficits.* Mol Endocrinol, 2016. **30**(4): p. 429-45.
58. Prince, V.E., R.M. Anderson, and G. Dalgin, *Zebrafish Pancreas Development and Regeneration: Fishing for Diabetes Therapies.* Curr Top Dev Biol, 2017. **124**: p. 235-276.
59. Chung, W.S., C.H. Shin, and D.Y. Stainier, *Bmp2 signaling regulates the hepatic versus pancreatic fate decision.* Dev Cell, 2008. **15**(5): p. 738-48.
60. Parsons, M.J., H. Pisharath, S. Yusuff, J.C. Moore, A.F. Siekmann, N. Lawson, and S.D. Leach, *Notch-responsive cells initiate the secondary transition in larval zebrafish pancreas.* Mech Dev, 2009. **126**(10): p. 898-912.
61. Lancman, J.J., N. Zvenigorodsky, K.P. Gates, D. Zhang, K. Solomon, R.K. Humphrey, T. Kuo, L. Setiawan, H. Verkade, Y.I. Chi, U.S. Jhala, C.V. Wright, D.Y. Stainier, and P.D. Dong, *Specification of hepatopancreas progenitors in zebrafish by hnf1ba and wnt2bb.* Development, 2013. **140**(13): p. 2669-79.
62. Flasse, L.C., J.L. Pirson, D.G. Stern, V. Von Berg, I. Manfroid, B. Peers, and M.L. Voz, *Ascl1b and Neurod1, instead of Neurog3, control pancreatic endocrine cell fate in zebrafish.* BMC Biology, 2013. **11**(1): p. 78.
63. Dalgin, G. and V.E. Prince, *Differential levels of Neurod establish zebrafish endocrine pancreas cell fates.* Developmental biology, 2015. **402**(1): p. 81-97.
64. Mohlke, K.L. and M. Boehnke, *Recent advances in understanding the genetic architecture of type 2 diabetes.* Hum Mol Genet, 2015. **24**(R1): p. R85-92.
65. Bonnefond, A. and P. Froguel, *Rare and Common Genetic Events in Type 2 Diabetes: What Should Biologists Know?* Cell Metabolism, 2015. **21**(3): p. 357-368.
66. Dziewulska, A., A.M. Dobosz, and A. Dobrzyn, *High-Throughput Approaches onto Uncover (Epi)Genomic Architecture of Type 2 Diabetes.* Genes (Basel), 2018. **9**(8).
67. Kleinberger, J.W. and T.I. Pollin, *Personalized medicine in diabetes mellitus: current opportunities and future prospects.* Ann N Y Acad Sci, 2015. **1346**(1): p. 45-56.
68. Roglic, G. and World Health Organization, *Global report on diabetes*. 2016, Geneva, Switzerland: World Health Organization. 86 pages.
69. Xue, A., Y. Wu, Z. Zhu, F. Zhang, K.E. Kemper, Z. Zheng, L. Yengo, L.R. Lloyd-Jones, J. Sidorenko, Y. Wu, Q.C. e, A.F. McRae, P.M. Visscher, J. Zeng, and J. Yang, *Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.* Nat Commun, 2018. **9**(1): p. 2941.
70. Leeder, S.R., *The history of insulin: the mystery of diabetes.* Med J Aust, 2013. **199**(4): p. 227.
71. Banting, F.G., C.H. Best, J.B. Collip, W.R. Campbell, and A.A. Fletcher, *Pancreatic Extracts in the Treatment of Diabetes Mellitus.* Can Med Assoc J, 1922. **12**(3): p. 141-6.
72. Zimmet, P.Z., D.J. Magliano, W.H. Herman, and J.E. Shaw, *Diabetes: a 21st century challenge.* Lancet Diabetes Endocrinol, 2014. **2**(1): p. 56-64.
73. Authors/Task Force, M., et al., *ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task*

Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur Heart J, 2013. **34**(39): p. 3035-87.

74. Kota, S.K., L.K. Meher, S. Jammula, S.K. Kota, and K.D. Modi, *Genetics of type 2 diabetes mellitus and other specific types of diabetes; its role in treatment modalities.* Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2012. **6**(1): p. 54-58.
75. Fajans, S.S., G.I. Bell, and K.S. Polonsky, *Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young.* N Engl J Med, 2001. **345**(13): p. 971-80.
76. Horikawa, Y., *Maturity-onset diabetes of the young as a model for elucidating the multifactorial origin of type 2 diabetes mellitus.* J Diabetes Investig, 2018. **9**(4): p. 704-712.
77. Antosik, K. and M. Borowiec, *Genetic Factors of Diabetes.* Arch Immunol Ther Exp (Warsz), 2016. **64**(Suppl 1): p. 157-160.
78. Amberger, J.S., C.A. Bocchini, F. Schiettecatte, A.F. Scott, and A. Hamosh, *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.
79. Wang, X., G. Strizich, Y. Hu, T. Wang, R.C. Kaplan, and Q. Qi, *Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction.* J Diabetes, 2016. **8**(1): p. 24-35.
80. Prasad, R.B. and L. Groop, *Precision medicine in type 2 diabetes.* Journal of Internal Medicine, 2019. **285**(1): p. 40-48.
81. Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
82. Khetan, S., R. Kursawe, A. Youn, N. Lawlor, A. Jillette, E.J. Marquez, D. Ucar, and M.L. Stitzel, *Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets.* Diabetes, 2018. **67**(11): p. 2466-2477.
83. Varshney, A., et al., *Genetic regulatory signatures underlying islet gene expression and type 2 diabetes.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(9): p. 2301-2306.
84. van de Bunt, M., J.E. Manning Fox, X. Dai, A. Barrett, C. Grey, L. Li, A.J. Bennett, P.R. Johnson, R.V. Rajotte, K.J. Gaulton, E.T. Dermitzakis, P.E. MacDonald, M.I. McCarthy, and A.L. Gloyn, *Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors.* PLoS genetics, 2015. **11**(12): p. e1005694-e1005694.
85. Kondratyeva, L.G., D.A. Didych, I.P. Chernov, E.P. Kopantzev, E.A. Stukacheva, T.V. Vinogradova, and E.D. Sverdlov, *Dependence of expression of regulatory master genes of embryonic development in pancreatic cancer cells on the intracellular concentration of the master regulator PDX1.* Dokl Biochem Biophys, 2017. **475**(1): p. 259-263.
86. Schwitzgebel, V.M., A. Mamin, T. Brun, B. Ritz-Laser, M. Zaiko, A. Maret, F.R. Jornayvaz, G.E. Theintz, O. Michielin, D. Melloul, and J. Philippe, *Agenesis of Human Pancreas due to Decreased Half-Life of Insulin Promoter Factor 1.* The Journal of Clinical Endocrinology & Metabolism, 2003. **88**(9): p. 4398-4406.
87. Wang, X., M. Sterr, I. Burtscher, S. Chen, A. Hieronimus, F. Machicao, H. Staiger, H.U. Haring, G. Lederer, T. Meitinger, F.M. Cernilogar, G. Schotta, M. Irmler, J. Beckers, M. Hrabe de Angelis, M. Ray, C.V.E. Wright, M. Bakhti, and H. Lickert,

Genome-wide analysis of PDX1 target genes in human pancreatic progenitors. Mol Metab, 2018. **9**: p. 57-68.
88. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
89. Milewski, W.M., S.J. Duguay, S.J. Chan, and D.F. Steiner, *Conservation of PDX-1 structure, function, and expression in zebrafish.* Endocrinology, 1998. **139**(3): p. 1440-9.
90. Francis, J., D.A. Babu, T.G. Deering, S.K. Chakrabarti, J.C. Garmey, C. Evans-Molina, D.G. Taylor, and R.G. Mirmira, *Role of chromatin accessibility in the occupancy and transcription of the insulin gene by the pancreatic and duodenal homeobox factor 1.* Mol Endocrinol, 2006. **20**(12): p. 3133-45.
91. Liberzon, A., G. Ridner, and M.D. Walker, *Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1.* Nucleic Acids Res, 2004. **32**(1): p. 54-64.
92. Kuzmich, A.I., D.V. Tyulkina, T.V. Vinogradova, and E.D. Sverdlov, *[Pioneer Transcription Factors in Normal Development and in Carcinogenesis].* Bioorg Khim, 2015. **41**(6): p. 636-43.
93. Kimmel, R.A., S. Dobler, N. Schmitner, T. Walsen, J. Freudenblum, and D. Meyer, *Diabetic pdx1-mutant zebrafish show conserved responses to nutrient overload and anti-glycemic treatment.* Sci Rep, 2015. **5**: p. 14241.
94. Wang, Y.J., J.T. Park, M.J. Parsons, and S.D. Leach, *Fate mapping of ptf1a-expressing cells during pancreatic organogenesis and regeneration in zebrafish.* Dev Dyn, 2015. **244**(6): p. 724-35.
95. Van Velkinburgh, J.C., S.E. Samaras, K. Gerrish, I. Artner, and R. Stein, *Interactions between areas I and II direct pdx-1 expression specifically to islet cell types of the mature and developing pancreas.* J Biol Chem, 2005. **280**(46): p. 38438-44.
96. Babu, D.A., T.G. Deering, and R.G. Mirmira, *A feat of metabolic proportions: Pdx1 orchestrates islet development and function in the maintenance of glucose homeostasis.* Mol Genet Metab, 2007. **92**(1-2): p. 43-55.
97. Gannon, M., L.W. Gamer, and C.V. Wright, *Regulatory regions driving developmental and tissue-specific expression of the essential pancreatic gene pdx1.* Dev Biol, 2001. **238**(1): p. 185-201.
98. Gerrish, K., J.C. Van Velkinburgh, and R. Stein, *Conserved transcriptional regulatory domains of the pdx-1 gene.* Mol Endocrinol, 2004. **18**(3): p. 533-48.
99. Gao, Y., J. Miyazaki, and G.W. Hart, *The transcription factor PDX-1 is post-translationally modified by O-linked N-acetylglucosamine and this modification is correlated with its DNA binding activity and insulin secretion in min6 beta-cells.* Arch Biochem Biophys, 2003. **415**(2): p. 155-63.
100. Kishi, A., T. Nakamura, Y. Nishio, H. Maegawa, and A. Kashiwagi, *Sumoylation of Pdx1 is associated with its nuclear localization and insulin gene activation.* Am J Physiol Endocrinol Metab, 2003. **284**(4): p. E830-40.
101. Lebrun, P., M.R. Montminy, and E. Van Obberghen, *Regulation of the pancreatic duodenal homeobox-1 protein by DNA-dependent protein kinase.* J Biol Chem, 2005. **280**(46): p. 38203-10.
102. McKenna, B., M. Guo, A. Reynolds, M. Hara, and R. Stein, *Dynamic recruitment of functionally distinct Swi/Snf chromatin remodeling complexes modulates Pdx1 activity in islet β cells.* Cell reports, 2015. **10**(12): p. 2032-2042.
103. Longo, A., G.P. Guanga, and R.B. Rose, *Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain.* Biochemistry, 2007. **46**(11): p. 2948-57.

104. Howe, D.G., et al., *ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics.* Nucleic Acids Res, 2013. **41**(Database issue): p. D854-60.
105. Haeussler, M., A.S. Zweig, C. Tyner, M.L. Speir, K.R. Rosenbloom, B.J. Raney, C.M. Lee, B.T. Lee, A.S. Hinrichs, J.N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G.P. Barber, D. Haussler, R.M. Kuhn, and W.J. Kent, *The UCSC Genome Browser database: 2019 update.* Nucleic Acids Res, 2019. **47**(D1): p. D853-D858.
106. Kwan, K.M., E. Fujimoto, C. Grabher, B.D. Mangum, M.E. Hardy, D.S. Campbell, J.M. Parant, H.J. Yost, J.P. Kanki, and C.B. Chien, *The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs.* Dev Dyn, 2007. **236**(11): p. 3088-99.
107. de la Calle-Mustienes, E., C.G. Feijoo, M. Manzanares, J.J. Tena, E. Rodriguez-Seguel, A. Letizia, M.L. Allende, and J.L. Gomez-Skarmeta, *A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts.* Genome Res, 2005. **15**(8): p. 1061-72.
108. Rueden, C.T., J. Schindelin, M.C. Hiner, B.E. DeZonia, A.E. Walter, E.T. Arena, and K.W. Eliceiri, *ImageJ2: ImageJ for the next generation of scientific image data.* BMC Bioinformatics, 2017. **18**(1): p. 529.
109. Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.Y. Tinevez, D.J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, *Fiji: an open-source platform for biological-image analysis.* Nat Methods, 2012. **9**(7): p. 676-82.
110. Warming, S., N. Costantino, D.L. Court, N.A. Jenkins, and N.G. Copeland, *Simple and highly efficient BAC recombineering using galK selection.* Nucleic acids research, 2005. **33**(4): p. e36-e36.
111. Kettleborough, R.N., et al., *A systematic genome-wide analysis of zebrafish protein-coding gene function.* Nature, 2013. **496**(7446): p. 494-7.
112. Postlethwait, J., A. Amores, A. Force, and Y.-L. Yan, *Chapter 8 The Zebrafish Genome*, in *Methods in Cell Biology*, H.W. Detrich, M. Westerfield, and L.I. Zon, Editors. 1998, Academic Press. p. 149-163.
113. Cox, A.R. and J.A. Kushner, *Area IV Knockout Reveals How Pdx1 Is Regulated in Postnatal β-Cell Development.* Diabetes, 2017. **66**(11): p. 2738-2740.
114. Yang, B.T., T.A. Dayeh, P.A. Volkov, C.L. Kirkpatrick, S. Malmgren, X. Jing, E. Renstrom, C.B. Wollheim, M.D. Nitert, and C. Ling, *Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with type 2 diabetes.* Mol Endocrinol, 2012. **26**(7): p. 1203-12.
115. Akerman, I., et al., *Human Pancreatic beta Cell lncRNAs Control Cell-Specific Regulatory Networks.* Cell Metab, 2017. **25**(2): p. 400-411.
116. Miguel-Escalada, I., et al., *Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes.* Nat Genet, 2019. **51**(7): p. 1137-1148.
117. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2007. **35**(Database issue): p. D5-12.
118. Binot, A.C., I. Manfroid, L. Flasse, M. Winandy, P. Motte, J.A. Martial, B. Peers, and M.L. Voz, *Nkx6.1 and nkx6.2 regulate alpha- and beta-cell formation in zebrafish by acting on pancreatic endocrine progenitor cells.* Dev Biol, 2010. **340**(2): p. 397-407.
119. Chung, J.H., M. Whiteley, and G. Felsenfeld, *A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila.* Cell, 1993. **74**(3): p. 505-14.

120. Mierzejewska, K., W. Siwek, H. Czapinska, M. Kaus-Drobek, M. Radlinska, K. Skowronek, J.M. Bujnicki, M. Dadlez, and M. Bochtler, *Structural basis of the methylation specificity of R.DpnI.* Nucleic acids research, 2014. **42**(13): p. 8745-8754.
121. Dobrin, R., D.M. Greenawalt, G. Hu, D.M. Kemp, L.M. Kaplan, E.E. Schadt, and V. Emilsson, *Dissecting cis regulation of gene expression in human metabolic tissues.* PLoS One, 2011. **6**(8): p. e23480.
122. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
123. van de Bunt, M., J.E. Manning Fox, X. Dai, A. Barrett, C. Grey, L. Li, A.J. Bennett, P.R. Johnson, R.V. Rajotte, K.J. Gaulton, E.T. Dermitzakis, P.E. MacDonald, M.I. McCarthy, and A.L. Gloyn, *Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors.* PLoS Genet, 2015. **11**(12): p. e1005694.

Supplementary Data

Supplementary Table 1. List of 126 candidate SNPs potentially associated to the development of diabetes. The list enumerates: Fasting glycemia- and T2D-associated SNPs retrieved from GWAS datasets; genes located nearby those SNPs; chromosomic localization, gene coding sequences (CDS) start and end coordinates (GRCh37/hg19), length of CDS and bibliographic references for each SNP.

Variant	Nearby gene	Chromosomic localization (GRCh37/hg19)	CDS start (GRCh37/hg19)	CDS end (GRCh37/hg19)	Length of CDS (bps)	References
rs35720761	*THADA*	Chr 2	43457975	43823185	365211	Mahajan et al 2018
rs7578597						Mohlke, K.L. et al. 2015
rs74382177						Pasquali et al. 2014
rs78681698						Pasquali et al. 2014
rs79629200						Pasquali et al. 2014
rs74469180						Pasquali et al. 2014
rs28700209						Pasquali et al. 2014
rs7567685						Pasquali et al. 2014
rs7581586						Pasquali et al. 2014
rs6752448						Pasquali et al. 2014
rs76282560						Pasquali et al. 2014
rs6728106						Pasquali et al. 2014
rs17039133						Pasquali et al. 2014
rs7600657						Pasquali et al. 2014
rs998768						Pasquali et al. 2014
rs74821293						Pasquali et al. 2014
rs72865297						Pasquali et al. 2014
rs74600494						Pasquali et al. 2014
rs11936387	*MAEA*	Chr 4	1283672	1333925	50254	Pasquali et al. 2014
rs6815464						Mohlke, K.L. et al. 2015
rs4689388	*WFS1*	Chr 4	6271577	6304992	33416	Pasquali et al. 2014
rs4320200						Pasquali et al. 2014
rs13107806						Pasquali et al. 2014
rs13127445						Pasquali et al. 2014

rs4273545						Pasquali et al. 2014
rs6830765						Pasquali et al. 2014
rs4457053	*ZBED3*	Chr 5	76372532	76383030	10499	Mohlke, K.L. et al. 2015
rs4457054						Pasquali et al. 2014
rs7708285						Pasquali et al. 2014
rs7732130						Pasquali et al. 2014
rs10440833	*CDKAL1*	Chr 6	20534457	21232404	697948	Mohlke, K.L. et al. 2015
rs9348441						Pasquali et al. 2014
rs9470794	*ZFAND3*	Chr 6	*37819499*	*38154624*	*335126*	Mohlke, K.L. et al. 2015
rs77114369						Pasquali et al. 2014
rs58692659						Pasquali et al. 2014
rs61332486						Pasquali et al. 2014
rs57995712						Pasquali et al. 2014
rs2908286	*GCK*	Chr 7	44183870	44229022	45153	Pasquali et al. 2014
rs4607517						Pasquali et al. 2014
rs515071	*ANK1*	Chr 8	*41510744*	*41754280*	*243537*	Mohlke, K.L. et al. 2015
rs508419						Pasquali et al. 2014
rs9694034						Pasquali et al. 2014
rs6989203						Pasquali et al. 2014
rs3802315						Sun W. et al 2018
rs516946						Sun W. et al 2018
rs750625						Sun W. et al 2018
rs11774700	*SLC30A8*	Chr 8	117962512	118188953	226442	Pasquali et al. 2014
rs13266634						Mohlke, K.L. et al. 2015
rs7041847	*GLIS3*	Chr 9	*3824127*	*4348392*	*524266*	Mohlke, K.L. et al. 2015
rs4237150						Pasquali et al. 2014
rs10814915						Pasquali et al. 2014
rs6476842						Pasquali et al. 2014
rs10814916						Pasquali et al. 2014
rs10811661	*CDKN2A*	Chr 9	*21967751*	*21994490*	*26740*	Pasquali et al. 2014

rs10811660						Pasquali et al. 2014
rs703977	*ZMIZ1*	Chr 10	80828792	81076285	247494	Pasquali et al. 2014
rs12571751						Mohlke, K.L. et al. 2015
rs7903146	*TCF7L2*	Chr 10	114710009	114927436	217428	Pasquali et al. 2014
rs231361	*KCNQ1*	Chr 11	2466221	2870340	404120	Pasquali et al. 2014
rs2237892						Mohlke, K.L. et al. 2015
rs3862791	*ARAP1*	Chr 11	72396114	72463434	67321	Pasquali et al. 2014
rs11603334						Mohlke, K.L. et al. 2015
rs15522243						Mohlke, K.L. et al. 2015
rs1552224	*ARAP1 /CENTD2*	Chr 11	72685069	72793599	108531	Mohlke, K.L. et al. 2015
rs56200889						Mahajan et al 2018
rs10842994	*KLHDC5 / KLHL42*	Chr 12	*27933187*	*27955973*	*22787*	Mohlke, K.L. et al. 2015
rs1127787						Mahajan et al 2018
rs12581729						Pasquali et al. 2014
rs10842991						Pasquali et al. 2014
rs10771372						Pasquali et al. 2014
rs3751239						Pasquali et al. 2014
rs10842992						Pasquali et al. 2014
rs10842993						Pasquali et al. 2014
rs11049161						Pasquali et al. 2014
rs7163757	*C2CD4A*	Chr 15	62359176	62363116	3941	Pasquali et al. 2014
rs7172432						Mohlke, K.L. et al. 2015
rs11634397	*ZFAND6*	Chr 15	*80351910*	*80430735*	*78826*	Mohlke, K.L. et al. 2015
rs1357335						Pasquali et al. 2014
rs1357336						Pasquali et al. 2014
rs7202877	*BCAR1*	Chr 16	75228187	75268053	39867	Mohlke, K.L. et al. 2015
rs72804106						Pasquali et al. 2014
rs8108269	*GIPR*	Chr 19	*46171502*	*46185717*	*14216*	Mohlke, K.L. et al. 2015
rs1800437						Mahajan et al 2018
rs11670462						Pasquali et al. 2014

rs55872740						Pasquali et al. 2014
rs10403962						Pasquali et al. 2014
rs10404142						Pasquali et al. 2014
rs10404527						Pasquali et al. 2014
rs10409882						Pasquali et al. 2014
rs8104845						Pasquali et al. 2014
rs2191349						Mohlke, K.L. et al. 2015
rs10244051						Pasquali et al. 2014
rs10950550						Pasquali et al. 2014
rs10228456						Pasquali et al. 2014
rs10228561	*DGKB*	Chr 7	*14184674*	*14881075*	*696402*	Pasquali et al. 2014
rs10228796						Pasquali et al. 2014
rs10258074						Pasquali et al. 2014
rs2191348						Pasquali et al. 2014
rs2191349						Pasquali et al. 2014
rs72695654	*ACSL1*	Chr 4	185676749	185747215	70467	Pasquali et al. 2014
rs735949						Pasquali et al. 2014
rs13431652	*G6PC2*	Chr 2	169757750	169766510	8761	Pasquali et al. 2014
rs4625	*AMT*	Chr 3	49454211	49460111	5901	Pasquali et al. 2014
rs1905506						Pasquali et al. 2014
rs1905504						Pasquali et al. 2014
rs7635100						Pasquali et al. 2014
rs7635470						Pasquali et al. 2014
rs11923694						Pasquali et al. 2014
rs11920090	*SLC2A2*	Chr 3	170714137	170744768	30632	Pasquali et al. 2014
rs11924648						Pasquali et al. 2014
rs61169219						Pasquali et al. 2014
rs7638998						Pasquali et al. 2014
rs5393						Pasquali et al. 2014
rs56198733	*PCSK1*	Chr 5	95726040	95768985	42946	Pasquali et al. 2014

rs4869273						Pasquali et al. 2014
rs12186664						Pasquali et al. 2014
rs17085593						Pasquali et al. 2014
rs59139497						Pasquali et al. 2014
rs10440833						Mohlke, K.L. et al. 2015
rs9348441	*CDKAL1*	Chr 6	*20534457*	*21232404*	*697948*	Pasquali et al. 2014
rs9348441						Pasquali et al. 2014
rs7945565						Pasquali et al. 2014
rs7945689	*CRY2*	Chr 11	45868957	45904799	35843	Pasquali et al. 2014
rs1401419						Pasquali et al. 2014
rs10501320	*MADD*	Chr11	47290927	47351582	60656	Pasquali et al. 2014
rs3783346	*WARS*	Chr 14	100800125	100842680	42556	Pasquali et al. 2014